MAIA

ERASMUS MUNDUS

JOINT MASTER IN MEDICAL IMAGING AND APPLICATIONS

Joint Master in Medical Imaging and Applications

Master Thesis Proceedings

Promotion 2021-23

www.maiamaster.org

MAIA

An international programme by the University of Girona (Spain), the University of Bourgogne (France) and the University of Cassino (Italy) funded by Erasmus + Programme.

Erasmus+ Universitat de Girona uB Université de Bourgogne

Published by the MAIA Master

www.maiamaster.org

This document is a compendium of the master thesis works developed by the students of the Joint Master Degree in Medical Imaging and Applications. Therefore, each work is independent on the other, and you should cite it individually as the final master degree report of the first author of each paper (Student name; title of the report; MAIA MSc Thesis; 2023).

Editorial

Computer aided applications for early detection and diagnosis, histopathological image analysis, treatment planning and monitoring, as well as robotised and guided surgery will positively impact health care during the new few years. The scientific community needs of prepared entrepreneurships with a proper ground to tackle these topics. The Joint Master Degree in Medical Imaging and Applications (MAIA) was born with the aim to fill this gap, offering highly skilled professionals with a depth knowledge on computer science, artificial intelligence, computer vision, medical robotics, and transversal topics.

The MAIA master is a two-years joint master degree (120 ECTS) between the Université de Bourgogne (uB, France), the Università degli studi di Cassino e del Lazio Meridionale (UNICLAM, Italy), and the Universitat de Girona (UdG, Spain), being the latter the coordinating institution. The program is supported by associate partners, that help in the sustainability of the program, not necessarily in economical terms, but in contributing in the design of the master, offering master thesis or internships, and expanding the visibility of the master. Moreover, the program is recognised by the European Commission for its academic excellence and is included in the list of Erasmus Mundus Joint Master Degrees under the Erasmus+ programme.

This document shows the outcome of the master tesis research developed by the MAIA students during the last semester, where they put their learnt knowledge in practice for solving different problems related with medical imaging. This include fully automatic anatomical structures segmentation, abnormality detection algorithms in different imaging modalities, biomechanical modelling, development of applications to be clinically usable, or practical components for integration into clinical workflows. We sincerely think that this document aims at further enhancing the dissemination of information about the quality of the master and may be of interest to the scientific community and foster networking opportunities amongst MAIA partners.

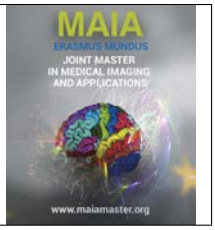
We finally want to thank and congratulate all the students for their effort done during this last semester of the Joint Master Degree in Medical Imaging and Applications.

MAIA Master Academic and Administrative Board

Contents

Medical Imaging and Applications

Master Thesis, June 2023



Temporal Image Registration using deep learning for 3D Fetal Echocardiography

Kazi Saeed Alam, Md Kamrul Hasan, Dr Choon Hwai Yap

Department of Bioengineering, Imperial College London, UK

Abstract

The fetal heart can experience congenital heart malformation and functional abnormalities. Ultrasound imaging plays a vital role in assessing the heart structure and function of the developing fetus due to its non-incisive nature. However, the detection of such abnormalities via mass screening is only 50%, suggesting a need for further improvement. Many researchers have been working in order to detect abnormalities in the heart from ultrasound imaging through segmenting cardiac chambers, valves, and blood flow patterns but most of the works are based on adult hearts. This motivates us to explore fetal echocardiographic images for which we collected 4D volume fetal heart images to perform temporal registration to segment the myocardium and left ventricle chamber from these images. Having a deep learning-enabled standardized approach to evaluation can improve precision and accuracy. Thus, in this project, we propose to develop methods for automatic 3D segmentation based on temporal registration from 4D fetal echo images. The 4D fetal echo images were collected and properly annotated with the help of an existing cardiac motion estimation algorithm. Our proposed model is built upon UNET based image registration model as a baseline with the residual branch, which is guided by a variational autoencoder to enforce structural features of the heart via latent space training and adversarial learning. We also plan to make the proposed model perform multi-scale registration. We have developed and tested our proposed network for both 2D (Adult images from CAMUS Dataset) and 3D (Fetal Data) segmentation which showed significant performance in both cases. As evaluation metrics, Mean squared error, and Dice Metric were computed both before and after the registration process.

Keywords: Fetal Echocardiography, Ultrasound, Image Registration, Variational Autoencoder, Adversarial Learning

1. Introduction

Ultrasound is one of the major imaging techniques that play a vital role to monitor cardiac functions and abnormalities. Due to its non-invasive nature ultrasound imaging has gained much popularity and has been used to assess heart structure and function by monitoring cardiac chambers, valves, and blood flow patterns. This imaging modality enables clinicians to diagnose and monitor congenital heart defects, providing valuable information for early intervention and management. Ciancarella et al. (2020), Sachdeva and Gupta (2020) showed the significance of the use of ultrasound in the field of cardiac imaging.

Heart structure and shapes such as Cardiac chambers, valves, blood flow patterns, etc can be used as good identifiers to detect and evaluate several cardiac diseases like Congenital heart defects, Coronary artery disease, Valvular heart diseases, Cardiomyopathies, etc. Research works from Green et al. (2023), Ong et al. (2020) shows that the information gained from the shape of the myocardium and heart chambers can give valuable insight which can detect and evaluate various Congenital heart defects. Researchers have been working to improve the detection of these diseases by employing automatic detection of heart structures and shapes. Having a deep learning-enabled standardized approach to automatically segment and detect can improve precision and accuracy.

Although most of the works are based on assessing the adult heart, the fetal heart also can experience congenital heart malformation and functional abnormalities. However, the detection of such problems via mass screening is only around 50%, suggesting a need for improvement. Being surprised at birth with fetal heart abnormalities instead of detecting them during mid-

gestation reduces the time available for planning and executing surgical treatment, and leads to poorer outcomes. Further, the evaluation approach of evaluating fetal heart health via fetal echo depends on many manual processes and involves subjective interpretation.

In our work, we are proposing a 3D temporal image registration-based segmentation technique to automatically detect and assess the left ventricle heart chamber and myocardium. The novelty of this research is mainly:

- A whole new 4D fetal echocardiography dataset with annotated 3D LV and myocardium masks for each 3D volume image. There is less research on fetal heart echocardiography due to the scarcity of well-produced datasets. We proposed an efficient workflow to manually segment the heart chamber and myocardium with temporal registration. An existing cardiac motion estimation algorithm has been used to assist the algorithm development. We hope that the publication of this new dataset will create a benchmark for further fetal heart echocardiography analysis and assessment.

- As the estimation of the deformation field by registration between two time points can help share the information between two segmentation branches, we are proposing a robust and efficient technique for temporal image registration for 4D fetal echocardiogram image volume.

We have proposed a Multi-class Anatomically Constrained and Multi-scale Registration (MACMR) framework in our research. The proposed registration method has the following integral parts:

- **Vanilla-DLIR**: The baseline architecture of the temporal registration is based on the typical UNET-like structure for performing image segmentation. The use of residual blocks helps to avoid the degradation of the features' quality as a non-zero regularizing path will skip over them. We are calling this baseline model Vanilla DLIR (Deep learning based image registration) as here the encoders of UNET try to extract features from lower to higher space and pass to the bottleneck whereas the task of the decoder is to produce the deformation field for the moving image so that it can be warped to match as much as possible as the target image.

- **AC-DLIR**: We have proposed to include a Variational encoder to enforce structural features of the heart via latent space training. The local segmentation-aware loss (fixed and moved labels) uses pixel-level predictions and may not ensure a satisfactory global match between the warped source and target anatomical masks. For this reason, the global latent space features can be beneficial for the network to perform better.

- **AdvAC-DLIR**: Moreover, we also propose to include adversarial learning as like zero-sum game theory (one agent's gain is another agent's loss), where the discriminator is used to classify moved and fixed images.

Still, there is room for performance improvement. Hence, we proposed Multi-class Anatomically Constrained and Multi-scale Registration (MACMR) framework. Additionally, we need a registration framework that can provide a suitable deformation field for all the scales of decoders in the proposed segmentation network to share the motion information. We have evaluated our proposed model for 2D as well as 3D volume datasets. For 2D data, the CAMUS 2D adult Echocardiography data were used from Leclerc et al. (2019) whereas for 3D data, the proposed fetal dataset. In order to validate our proposed framework, we have conducted several experiments on the existing DL-based registration pipeline.

2. Literature Review

Researchers have worked in the field of medical image registration in various directions. A broad topic like image registration can be classified into various objectives. The methods can be interpatient or intra-patient (same patient at different time points), and the images can be from one single imaging technique (unimodal) or can be of multimodal techniques. The registration methods can be deformable, affine, or simply rigid. Also based on the organ of interest, it can be the brain, lungs, heart, or even tumors and so on. Input images can be of different types of dimensions or combinations of them. In our work, we have tried to cover the unimodal, intrapatient fetal echocardiographic registration based on 4D volume images. In the following sections, the recent trends in image registration as well as focus based on ultrasound techniques will be explored.

2.1. Deep learning based Image Registration

We will restrict the discussion of trends in medical image registration in deep learning-based (DL) techniques as the recent works have shown an upward trend in the domain of image registration yielding state-of-the-art for various applications. The use of conventional similarity-based metrics such as mean-squared error, structural similarity, cross-correlation, mutual information, etc work well for unimodal image registration in the case of CT or MRI images, as shown in Gong et al. (2017), Heinrich et al. (2012). But the presence of noise such as in ultrasound images or in the case

of multi-modal registration they failed to perform satisfactorily (Rivaz et al. (2014)). Many researchers have replaced these conventional methods with CNN-based deep-learning image registration and achieved success.

Cheng et al. (2018) proposed an unsupervised learning-based registration method to train a classifier to learn the deformation field using continuous probabilistic values for similarity measures. In their work, they have claimed the learned deep similarity metric outperforms MI as in conventional methods in brain T1-T2 registration. The challenge was to have a smooth first-order derivative to have a better overlap between the fixed and moving images. Other works from Simonovsky et al. (2016), Ferrante et al. (2018) also explored the use of deep similarity metrics with unsupervised or weakly supervised training. The challenge of these works was to acquire an accurately aligned image. Images with noise such as ultrasound or in the case of multi-modal image pairs, the same performance will be difficult to achieve.

Compared to other modalities, ultrasound images are a bit challenging due to the image acquisition technique and also due to the presence of artifacts such as speckle noise. Haskins et al. (2018) in their work showed the comparison of multimodal image registration based on deep learning similarities. They have shown the CT-MRI pairs have better registration than the MRI-US pair in the case of the use of single similarity metrics. Ferrante et al. (2018) proposed the use of multiple metrics instead of single ones and showed improved performance for ultrasound image registration.

Wu et al. (2016) have introduced the use of variational autoencoders to perform latent space training, they have shown the use of both local and global features improved the performance of training with only local features. They have used the segmented masks as well as intensity images of brain MRI to perform the latent space training for image registration. They used a stack of autoencoders for the model to learn the latent space features and compared the result with Dice Similarity Coefficient (DSC). Although the dice similarity between the masks improves the smoothness of the shapes of the human organ, the intensity similarity between fixed and moving image still needs further improvement.

To provide better regularization which was lacking in the works discussed by VAs, some researchers proposed the use of adversarial learning or GAN-based models. As human organs are highly regular, better regularization is needed to have plausible shapes in the produced output. Yan et al. (2018) in their work have trained GAN-based networks to discriminate between ground truth-based and prediction-based transform to deform images. In their work, they have used the adversarial loss to optimize the accurate transform to deform the fixed image. Fu et al. (2020) also showed similar improvement in registration performance by introducing adversarial loss. GAN-based models helped to generate more plausible and medically acceptable structures and shapes after registration in these research works. However, the similarity in intensity matching for GANs still needs to be investigated thoroughly. Some recent advances also show the use of transfer learning, LSTMs, one-shot predictions, Faster RCNN, etc in Xie et al. (2022), Fechter and Baltas (2019), Jaderberg et al. (2015). However, all these have been applied to mostly CT and MRI images. While applying ultrasound images, most of them do not show any satisfactory improvement.

2.1.1. Image Registration in EchoCardiography

For cardiac chamber segmentation, ultrasound images can be acquired in two chambers (A2C) or four chambers (A4C) view. An optical flow estimation-based technique for deep, fully convolutional networks was suggested by Jafari et al. (2018). Jafari et al. (2019) also proposed the use of semi-supervised learning where they have incorporated inverse mapping with the use of adversarial learning and inverse mapping of the moved and target masks for LV segmentation. Yoon et al. (2021) in their work showed the use of Regional-CNN to extract geometrical attributes to perform LV segmentation. Variational autoencoders as discussed before have been also used for cardiac chamber segmentation tasks in cardiac ultrasounds. Painchaud et al. (2019) used VAs to represent the latent space training.

Some works have been also done in fetal echocardiography. Yang et al. (2020) in their work have used DeeplabV3 with UNET to segment the left ventricle chambers for the fetus's heart. Dong et al. (2019) also worked with A4C view for residual visual block network-based segmentation of the fetal heart chamber. But as the dataset is very limited for fetal echocardiography, still the performance of these models needs to be investigated further.

3. Dataset Description

4D echocardiography dicom images were acquired for studying out of which 4 were healthy fetuses and the rest were diseased cases. The fetuses were of mixed gender and different ethnic groups (Chinese, Indian, Malay). Most of the cases had a gestation age between 22 to 32 weeks. The images were obtained in accordance with protocol 2014/00056 from Domain Specific Review Board and with the consent of all the participants. The 4D echo images were carried out with GE Voluson 730 ultrasound connected to the RAB 4-8L transducer (GE Healthcare Inc., Chicago, Illinois, USA) which has approximated 154 μm axial resolution and around 219 μm lateral resolution along with a transducer of 5 MHz.

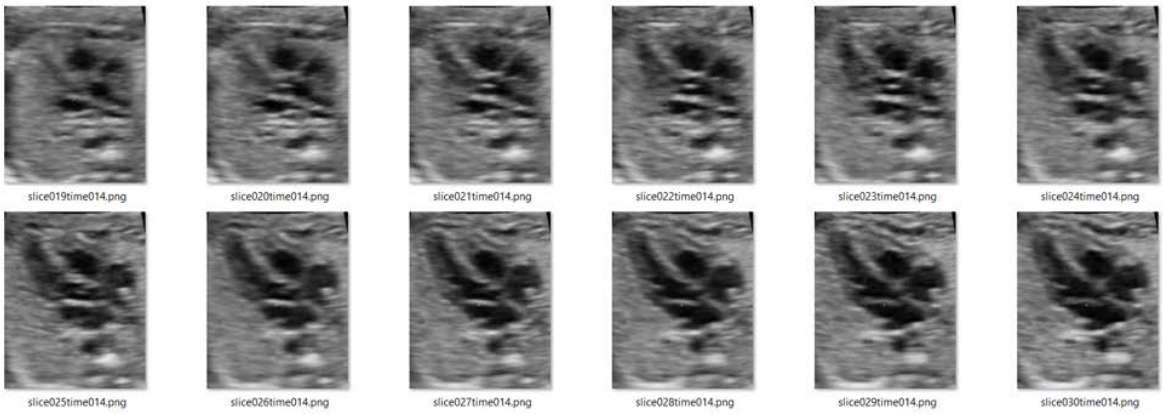


Figure 1: Visualization of intensity image slices (Patient001).

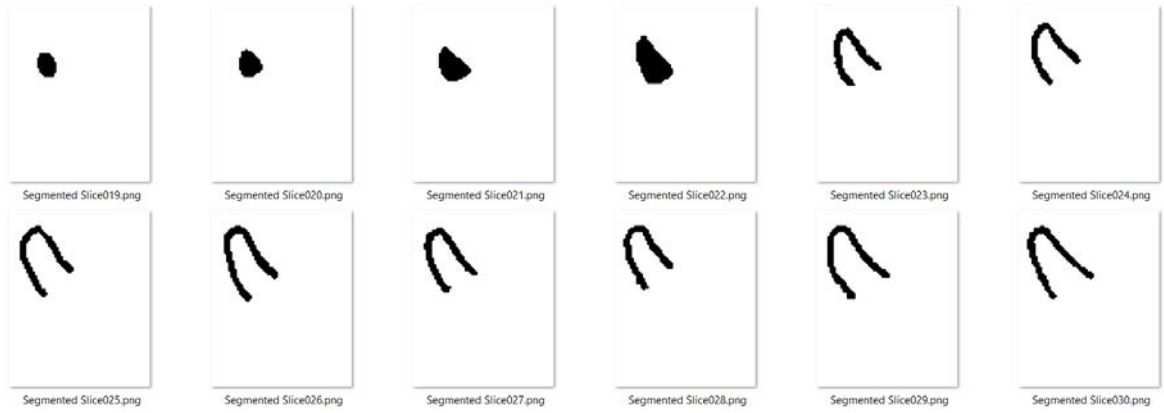


Figure 2: Visualization of annotated masks (Patient001).

3.1. Data Preparation

As the ultrasound raw images acquired were in 4D dicom format, they needed to be transformed to 3D format. This will help to extract the time points for each patient's case. To do so, a special software named "*4D View*" developed by GE Healthcare was used. This step will generate a cine sequence or cine loop video which will hold the information of desired slices for each time point. The inputs for each dicom series image were the cine length which means the number of cine sequences to be stored in the cine loop, start and end slices considering the proper visualization of the region of interest which is the left ventricle in this case, and the step size which means the distance (in millimeters) between each slice in the cine loop sequence. After that, the dicom series 4D images will be transformed into a series of cine sequence videos which will hold the temporal information for all the slices. After extracting time points for each case, the next step is to extract slices for each time point for all the patient cases. A Matlab script was written to extract the video frames from each video setting the distance between the vertical slices. The start and end slices were chosen and then the picture frame was cropped so that each slice will hold the region of interest and not contain unnecessary pixels. After extracting the slices from each time point they are ready for image registration to get the deformation field. A set of slices as an example after the data preparation step is demonstrated in Figure 1.

3.2. Registration

The target of this step is to register the slices with respect to each time point to derive the deformation field. Each slice image at a particular time point (t_n) will be registered with respect to the initial time point (t_0) and the previous time point image (t_{n-1}). For performing image registration, *SimpleElastix* by Lowekamp et al. (2016) and *Cardiac motion estimation* library by Wiputra et al. (2020) were used. Here, the cardiac motion tracking uses the Fourier b-splines spatiotemporal motion model to fit the deformation fields. It requires the initial and final time points with the number of slices to be specified. After setting up the paths and initializing the bspline-solver, it performs the pairwise registration and stores the displacement fields for each pair with the scaling and transformation parameters. For each single

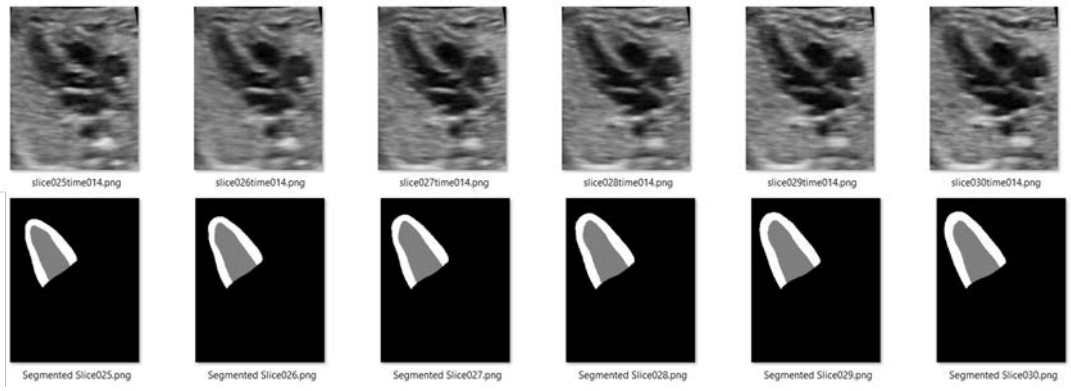


Figure 3: Sample intensity image slice and mask pairs.

image slice at (t_n), there will be two displacement fields which will be later combined using a weighted average to transform and derive the mask for that time point (t_n).

3.3. Segmentation

The next step after registration is manual segmentation or annotation. For this, two specific time points (preferably end-systolic/initial and end-diastolic/final) were chosen for each case. After that, all the slices for those time points were manually annotated. For performing the annotation, a quick and interactive segmentation technique called "*Lazy Snapping*" by Li et al. (2004) was chosen. It helps to choose the foreground and background by using a marker and based on that it generates the mask for each particular slice. The greater number of slices, the more robust data but also the number of slices might make the process a bit time-consuming as manual segmentation takes a considerable amount of time. After generating the segmentation mask, they were also checked and assessed by experts and their feedback was received. The generated segmentation masks would be irregular or not smooth enough as they might have staircase effects or holes. These will be corrected and smoothed in the later steps before generating the masks for other time points. A set of segmented masks as an example after the manual annotation using lazy snapping can be seen in Figure 2.

3.4. 3D Reconstruction

After generating the left ventricle masks for end-systolic/initial and end-diastolic/final time points, the next step is to combine these 2D slices to reconstruct the 3D mask for those points. For 3D reconstruction, "*VMTK (Vascular Modeling Toolkit)*" by Izzo et al. (2018) has been used which is a popular software for vascular image reconstruction and geometric analysis. The paths for all the slices were given as inputs and the result was the 3D reconstructed mask for the left ventricle at a given time point. As the results from the lazy snap step were not smooth and contained some artifacts, these 3D masks were corrected and smoothed with the help of an expert using "*Geomagic Wrap*" Software. This reverse engineering software helped to smoothen and regularize the mask by removing the artifacts. After the 3D mask was approved by the expert, later it was used to generate the other time point masks. To generate the 3D masks for other time points, the deformation fields obtained in the image registration step were used. Finally, for all the patient cases, 3D masks were generated for all the time points which were later used for training and testing the deep learning models. An example of the 3D mask can be seen in Figure 4.

3.5. Image Preprocessing

The intensity images acquired through the ultrasound scanner generated some artifacts like constant white boxes or arrows in the image which can be seen in Figure 5. For better performance during train, these constant regions should be removed or replaced by the neighboring pixel intensities as they might generate undesired results during training. As these artifacts were common and at the same position for all the images over slice and time for any cases, the same step for removing these artifacts from one image has been applied for all. To remove, the constant regions from the image, a simple linear interpolation method was used. In this method, an interpolation line was drawn between the left and the right pixel of the defected area, and then the defected area was interpolated using the intensity values from the interpolation line. The sample results can be seen in the same Figure 5.

3.6. Image-Mask Pair Generation

In the last step, the inner wall of the reconstructed masks was filled and reconstructed masks were binarised where (class 0 represents the background, 1 for the cavity of the left ventricle, and 2 for the myocardium. Later 3D masks were paired with the corresponding 3D intensity images for each time point to finalize the dataset for training. In the end, 14 4D echocardiography

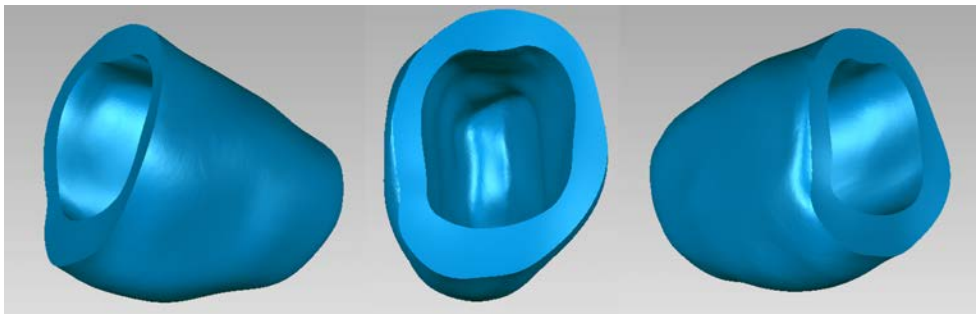


Figure 4: Example 3D Mask for Patient001 (Time014).

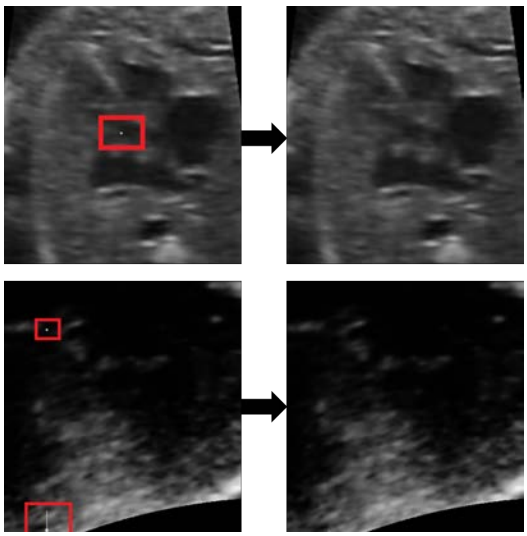


Figure 5: Intensity image artifact removal example.

images were transformed into a total of 518 3D images where each of the 3D images holds around 40 2D slices. As the nifty formatted files are hard to visualize in the report, a sample of slices for image and mask pairs are shown in Figure 3.

4. Experimental Methods

Let's assume f and m are two volume images that can be referred to as target/fixed image and moving image. The goal is to deform the moving image so that the anatomical location for all the voxels in fixed and moved images will be the same. Deep learning-based image registration (DLIR) neural networks will be used to model the displacement field which will transform all the voxels in the moving image so that they can be aligned with the fixed image. Let's say, the displacement field u will be modeled by CNN as the function $g_\theta(f, m) = \mathbf{d}$, where $\mathbf{d}$ is the displacement field and θ is the set of parameters learned by the CNN network. The main aim is to optimize the set of parameters so that the expected loss function can be minimized using Stochastic Gradient descent. Several approaches and experiments have been conducted to perform optimal image registration. The approaches will be discussed as follows.

4.1. Approach 1: Vanilla-DLIR

The underlying architecture of Vanilla DLIR is based on the traditional UNET architecture by Chen et al. (2021); Ronneberger et al. (2015) used for segmentation. The UNET consists of encoding and decoding layers with residual skip connections. This can be seen in Figure 6. The network used receives input fixed and moving images both of $256 * 256 * 32$ sizes which are concatenated to 2-channel 3D images. The 3D convolution is applied both in the encoding and decoding layers with a kernel size of 3, the stride is kept as 2 which is followed by Batch Normalization and ReLU layers. Max pooling is applied for downsampling in the encoding layers to reduce the spatial dimension of the image by half. The number of channels increases where the image size is reduced for the coarser representation of the input in the pyramid hierarchy. The bottleneck layer after the encoding layers captures the most abstract feature of the input image volume.

Then, the decoding layers perform the upsampling and convolutional operations to generate the displacement field. The convolutional layers consist of transposed 3D convolutions followed by batch normalization and ReLU layers. Skip connections from the encoding layers directly applied by concatenating. The conventional path cannot degrade the features' quality as a non-zero regularizing path will skip over them. On the other hand, the direct skipping of the non-zero regularizing path cannot hamper the performance as it has been added to the conventional path's learned features. Each layer of the decoding stage generates a finer spatial scaled image for generating the deformation field as an output of the final convolutional layer containing a 1*1 image filter and a softmax activation function.

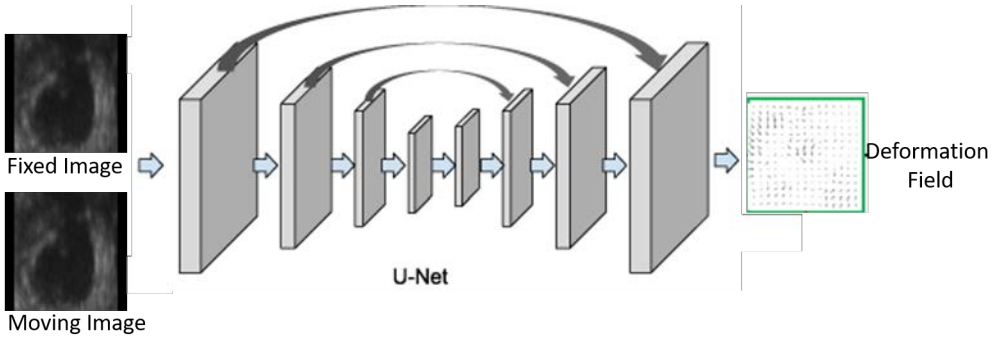


Figure 6: UNET for Image Registration with Skip connection.

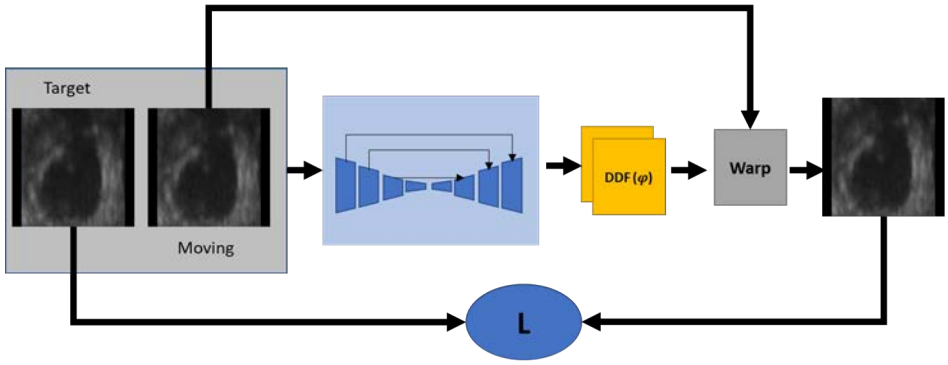


Figure 7: Vanilla-DLIR architecture.

4.1.1. *Vanilla-DLIR Loss Functions*

For vanilla-DLIR, unsupervised loss functions have been incorporated which consists of two components mainly. The first component is the similarity loss for having a better approximation of the fixed image in appearance for the moved image. Whereas, a regularization loss function called binding energy loss is used to penalize the non-regular spatial differences in order to have a smoother and more plausible displacement field. The equation for the total unsupervised loss is as follows where λ is a regularization parameter.

$$\mathcal{L}_{us}(f, m, d) = \mathcal{L}_{\text{sim}}(f, m \circ d) + \lambda \mathcal{L}_{\text{smooth}}(d) \quad (1)$$

There are a couple of similarity loss functions that can be used such as mean squared error(MSE), cross-correlation(CC), etc but for this work, Global Mutual Information(GMI) loss has been used. The statistical dependency or mutual information between two random variables, generally the fixed and the deformed moving image using the displacement field, is measured by the GMI loss. GMI loss seeks to maximize the similarity in appearance between the fixed image and the produced output. The model is compelled to acquire meaningful and instructive representations by maximizing mutual information. Firstly, the mutual information between the local patches is calculated and then the local patches MIs are aggregated to get the global mutual information. To calculate the mutual information for the patches in f and m, the following equation can be used,

$$I(f_x; m_y) = \sum_{x \in X} \sum_{y \in Y} P(f_x, m_y) \log_2 \left(\frac{P(f_x, m_y)}{P(f_x) P(m_y)} \right) \quad (2)$$

Higher mutual information yields a better alignment, so minimizing the negative GMI loss, the model tries to maximize the MI between the fixed and moved images.

GMI loss enforces the model to approximate the fixed image but the produced output may not be as smooth as desired. To have a smooth and more physically realistic deformed moving image, binding energy loss is also used in addition to GMI loss. Using a diffusion regularizer can leverage the spatial gradients of the deformation, u.

$$\mathcal{L}_{\text{smooth}}(d) = \sum_{\mathbf{d} \in D} \|\nabla \mathbf{u}(\mathbf{p})\|^2 \quad (3)$$

The differences between neighboring pixels in the 3D image are used to approximate the spatial gradient. The resulting architecture of Vanilla-DLIR with its loss function can be seen in Figure 7.

4.2. Approach 2: Anatomically Constrained DLIR

Anatomical masks of the Myocardium and left ventricle cavity are available from the data annotation part, the vanilla-DLIR can leverage from it. Balakrishnan et al. (2018) and Hu et al. (2017) in their respective research works showed that, the use of deformed segmentation masks during training enhances the performance of image registration in Vanilla-DLIR. In order to leverage the segmentation masks, first the registration field d, derived from the model network was used to deform the fixed image mask. After that, the segmented mask of the deformed image became available during training. As the segmented masks assign labels to the specific regions in the image, the same specific region in the fixed mask and deformed mask should also overlap. That was the key idea of getting the use of supervised loss in addition to the unsupervised loss for Vanilla-DLIR. Dice (1945) shows to quantify this volume overlap, Dice Score can be used. For example, the regions of either myocardium or left ventricle cavity, in this case, can be expressed in terms of the fixed and moved image can be expressed as $r_f^v and r_m^v \circ d$. The dice score can be computed to quantify the overlap of both regions as follows.

$$\text{Dice}\left(r_f^v, r_m^v \circ d\right) = 2 \cdot \frac{\left|r_f^v \cap (r_m^v \circ d)\right|}{\left|r_f^v\right| + |r_m^v \circ d|} \tag{4}$$

The dice score lies between 0 to 1, from no overlap to complete overlap. The dice score loss was defined $\mathcal{L}_{\text{dice}}$ over the whole segmented regions $v \in [1, V]$ as:

$$\mathcal{L}_{dice}\left(r_f, r_m \circ d\right) = -\frac{1}{K}\sum_{v=1}^{V} \text{Dice}\left(r_f^v, r_m^v \circ d\right) \tag{5}$$

4.2.1. Latent Space Consideration

In addition to dice score loss, the global anatomical constraint was also considered to compute the global loss. The local segmentation-aware loss computed by dice loss (fixed and moved labels) uses pixel-level predictions and may not ensure a satisfactory global match between the warped source and target anatomical masks shown by Oktay et al. (2017). Here, the segmentation masks for the fetal echo image volumes, represent the myocardium and left ventricle cavity. Segmentation masks represent pathological entities like brain tumors or skin lesions, which are very irregular in shape and topology. Whereas, Human organs like this scenario are highly regular, and are used to constrain registration. So, the plausibility of the shape is very important to get the correct registered images. For this reason, the latent space of the both target and the moved mask was considered to compute the global loss function. The global loss function considers the anatomical plausibility of the deformed source mask when comparing it to the target mask. Moreover, Oktay et al. (2017) also shows that, local dice loss acts at the pixel level, and back-propagated gradients are parametrized exclusively by pixel-wise individual probability components and provide little global context. To put global context in the loss computation, variational encoders were used to transform the target and moved masks to latent space, and compute global loss. The idea of a variational autoencoder can be understood in the next section and visualized in Figure 8.

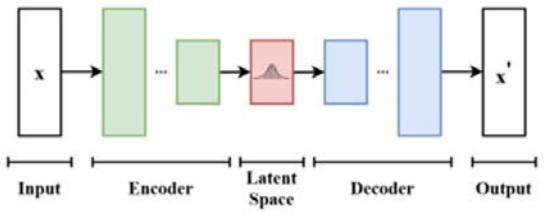


Figure 8: Learning global anatomical features.

4.2.2. Variational Autoencoder

To compute the global loss from the observations, the segmented masks needed to be transformed into latent space. A Variational autoencoder exactly does the same as shown by Oktay et al. (2017). Variational autoencoders(VAE) provide a probabilistic manner to describe the observations in latent space. In this work, the idea of VAE was adapted with a little bit of change in the architecture can be seen in Figure 9 to make it work for fetal echo masks for the myocardium and left ventricle. VAE contains two parts, encoders and decoders with a bottleneck layer. Encoders learn effective data encoding from datasets and pass it into bottleneck architectures. The autoencoder's decoder employs latent space in the bottleneck layer to generate dataset-like images. These results backpropagate from the neural network in the form of the loss function. For this work, the encoder part had 4 hierarchical stages each containing a block of convolutional neural network having kernel size=3 layer followed by batch normalization and ReLU layers. The downsampling was done by max pooling having kernel size=3, stride=2, and pooling=1. Residual connections were introduced at each stage to improve the flow of gradients during training. The inputs of the encoder were the single channel mask volumes of size $256 * 256 * 32$ which were halved at later stages. The bottleneck layer was a linear network transforming the output from the encoder to the latent space and passing it to the decoder. The decoder has the same 4 stages as the encoder where each stage has 3 blocks of convolutional neural network followed by batch normalization and ReLU.The upsampling was done with a scale factor

of 2 using trilinear interpolation. Finally, after the last stage, the top input-like images were reconstructed.

The loss functions for the variational autoencoders were a combination of 4 loss functions.

- Dice Score loss
- Euclidean L2 norm loss
- Structural Similarity loss
- Kullback-Leibler(KL) Loss

The dice score loss is computed between the input and reconstructed image using the equation 5.

The Euclidean L2 norm loss computes the Euclidean distance between the input images and the reconstructed ones. Let's say, if i and r are the input and reconstructed masks respectively, the L2 loss was computed by the following equation:

$$\mathcal{L}_{L2}(i,r) = \frac{1}{N}\sum_{n=1}^{N}(i_n - r_n)^2 \tag{6}$$

L2 norm penalizes the larger distances between the voxels in input and reconstructed masks more than the smaller distances.

To assess the quality of the image reconstruction by guiding the image generation, the structural similarity measure index was also computed as shown by Wang et al. (2004). Structural similarity loss can be computed to penalize the dissimilarity between the input and the reconstructed masks. The following equation was used to compute the SSIM loss:

$$\mathcal{L}_{SSIM}(i,r) = 1 - \frac{(2\mu_i\mu_r + C_1)(2\sigma_{ir} + C_2)}{\left(\mu_i^2 + \mu_r^2 + C_1\right)\left(\sigma_i^2 + \sigma_r^2 + C_2\right)} \tag{7}$$

where μ_i and μ_r are the average pixel intensities of i and r, σ_i and σ_r are the standard deviations of pixel intensities. Finally, σ_{ir} is the covariance between the pixel intensities of the two images. C1 and C2 are small constants added to stabilize the division when the denominator approaches zero.

The regularization loss named Kullback-Leibler (KL) divergence in Kingma and Welling (2014) forces the distributions returned by the encoder to be close to a standard normal distribution. KL loss will be a good representative to assess the discrepancy between the latent and desired distribution, and thus in generative models like VAEs, the KL divergence can be often used as a regularization term. The goal is to penalize the discrepancy between the learned latent distribution and a prior standard normal distribution. Let's say, for the standard normal distribution prior is $P(z)$, and the learned approximate posterior $Q(z|x)$, KL loss will be:

$$\mathcal{L}_{KL}(P(z), Q(z|x)) = \frac{1}{2}\sum\left(\mu^2 + \sigma^2 - \log(\sigma^2) - 1\right) \tag{8}$$

where μ and σ are the mean and standard deviation of the approximate posterior distribution $Q(z|x)$ for each latent variable z and will be summed for all latent variables. Finally, the variational autoencoder is trained to optimize the total loss function which can be described as:

$$\mathcal{L}_{va}(i,r,P(z),Q(z|x)) = \mathcal{L}_{dice}(i,r) + \mathcal{L}_{L2}(i,r) + \mathcal{L}_{SSIM}(i,r) + \mathcal{L}_{KL}(P(z),Q(z|x)) \tag{9}$$

For training and validating the variational autoencoder, out of 518 3D annotated volume masks discussed in the dataset description section, 452 volume masks were used for training and the rest for validation. To improve the generalization of VAE, some data augmentation techniques like flipping and center-cropping were also used. An example of the results after the training of VAE can be seen in Figure 11.

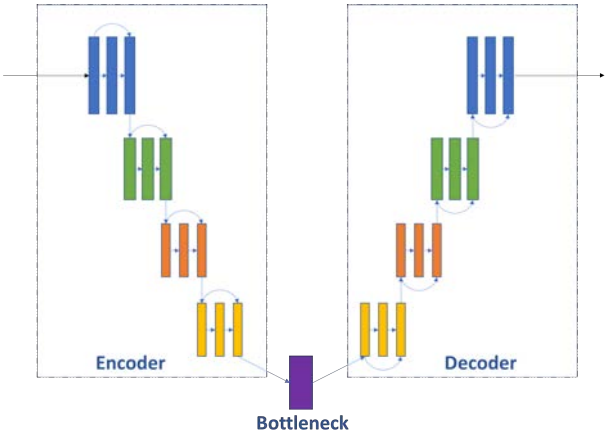


Figure 9: Variational Autoencoder Architecture.

4.2.3. AC-DLIR Loss Functions

The unsupervised loss introduced for vanilla-DLIR and the dice score loss from equation 5 are combined. In addition to that, image global loss is computed too. For computing global loss, the latent space consideration from VAE is used. Both the input and predicted mask are passed by the variational autoencoder model to generate the reconstructed masks. The global loss is computed between these two reconstructed masks both for the myocardium and left ventricle and added together. The global loss is the computation of the L2 norm which is discussed in equation 6. The total loss with anatomical constraint consideration for AC-DLIR is:

$$\begin{aligned}\mathcal{L}_a\left(f,m,r_f,r_m,d\right) &= \mathcal{L}_{us}(f,m,d)\\ +\beta\mathcal{L}_{dice}\left(r_f, r_m \circ d\right) &+ \gamma\mathcal{L}_{L2}(r_f, r_m)\end{aligned} \tag{10}$$

where, both β and γ are regularization parameters. Fi-

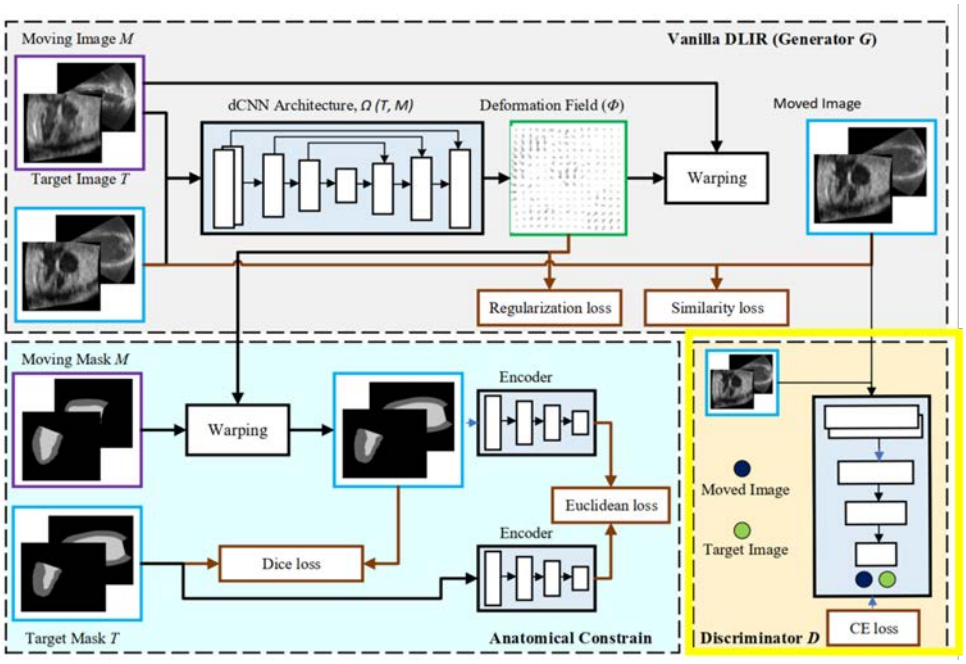


Figure 10: Proposed Adversarial Anatomically Constrained (AdvAC) DLIR architecture.

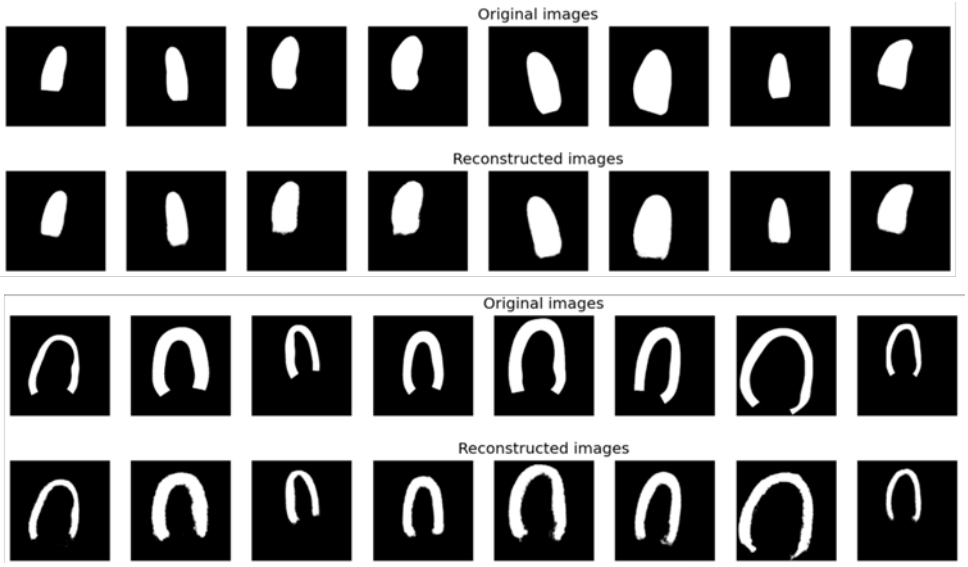


Figure 11: Original and reconstructed masks by VAE (top: left ventricle, bottom: myocardium).

nally, the architecture used for AC-DLIR can be visualized in Figure 10 excluding the highlighted part.

4.3. Approach 3: Adversarial AC-DLIR

The next addition to the network proposed is the inclusion of adversarial learning. As shown by Mahapatra et al. (2018), the use of the GAN network as a zero-sum game theory could be beneficial for learning deformable fields in image registration. In the proposed network, the part of AC-DLIR for generating the deformable images with the produced deformation field was treated as a generator for the adversarial network. In addition to that, a discriminator was also created which was able to classify the fixed and moved images. The architecture of the discriminator consists of 5 layers each containing convolutional blocks with 2 residual units outputting 8,16,32,64 and single channels respectively. The input was the single channel input image volume. Kernel size was kept at 3 with strides 2,2,2,2 and 1 at the respective layers and with LeakyReLU activation. The dropout layer was also used with a probability of 0.10. For the loss function of both the generator and discrim-

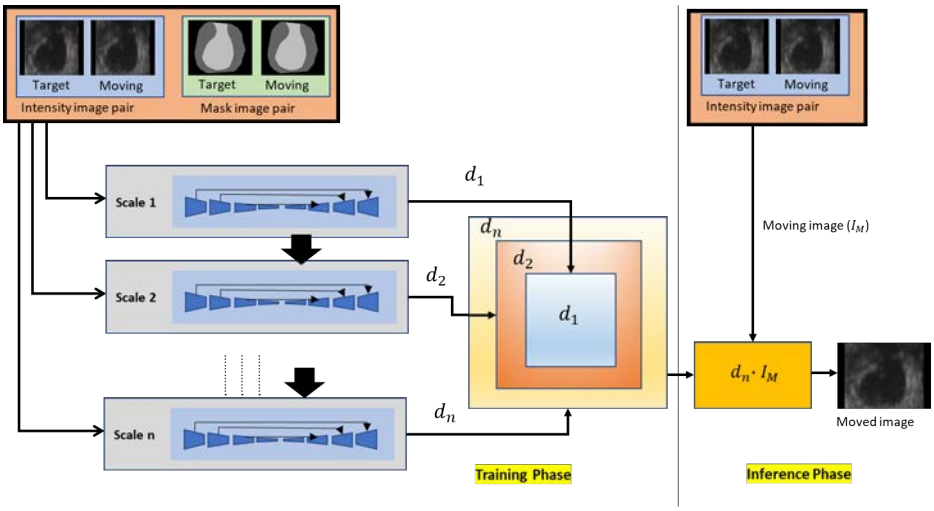


Figure 12: Proposed MACMR architecture.

Models	MSE	Dice Score			Mean Dice ± std
		BG	Myo	LV	
Without Registration	0.00972	0.96678	0.69391	0.76046	0.80235±0.05491
Vanilla-DLIR	0.0042	0.97352	0.74977	0.87523	0.88487±0.03261
AC-DLIR	0.00598	0.97972	0.81437	0.91935	0.90303±0.03447
Adv-DLIR	0.00533	0.97429	0.79278	0.86842	0.85733±0.04129
AdvAC-DLIR	**0.00589**	**0.98742**	**0.82751**	**0.93573**	**0.91689±0.02596**
MACMR	**0.00489**	**0.98779**	**0.84871**	**0.95423**	**0.94245±0.02474**

Table 1: Comparison of proposed registration models on CAMUS 2D Dataset.

Models	MSE	Dice Score			Mean Dice ± std
		BG	LV	Myo	
Without Registration	0.00377	0.99093	0.78917	0.72605	0.83539±0.12798
Vanilla-DLIR	0.00296	0.98699	0.70087	0.58543	0.75776±0.04036
AC-DLIR	0.00251	0.98959	0.73347	0.64435	0.80013±0.05401
Adv-DLIR	**0.00339**	**0.99031**	**0.73836**	**0.67389**	**0.80989±0.05142**
AdvAC-DLIR	**0.00258**	**0.99089**	**0.79884**	**0.73482**	**0.84668±0.04586**

Table 2: Comparison of proposed registration models on Fetal 3D Dataset.

inator, the binary cross-entropy loss was used. While training, the generator, and discriminator will fight o gain over each other as the task of the generator would be creating as much as plausible images as the fixed image whereas the discriminator would try to discriminate them. The loss from the generator was added to the $\mathcal{L}_a(f, m, r_f, r_m, d)$ from equation 10 as the deformable field generated by training would be capable of better generalization if the loss of the generator was being optimized.

$$\mathcal{L}_{adac}\left(f, m, r_f, r_m, d, s_m\right) = \mathcal{L}_{us}(f, m, d) + \beta\mathcal{L}_{dice}\left(r_f, r_m \circ d\right) + \gamma\mathcal{L}_{L2}(r_f, r_m) + \phi\mathcal{L}_g(m, s_m) \quad (11)$$

In this loss function, equation, ϕ is a regularization parameter set as 0.0001, m and s_m are the moved image and assigned real labels to the moved image. The final architecture after adding the adversarial network to the AC-DLIR can be seen in figure 10.

4.4. Approach 4: Multi-Scale Registration (MACMR)

The final proposal to improve the performance of image registration is Multi-scale (multi-resolution) training, where trained parameters on the lower scale will be used to initialize the higher-scale training. As the features learned at the lower scales can guide the training for the higher scale, the network at a higher scale will have a better initialization. Better initialization of the network should result help the network converge faster to achieve better performance. Moreover, it can be shown that Multi-resolution training helps the network to learn both local and global information. It can improve the performance of the model with various

scales and enhance its overall performance. The proposed MACMR architecture is demonstrated in Figure 12.

5. Results

We have implemented all the methods discussed above in Pytorch. The experiments were performed on NVIDIA GeForce RTX 3090 Ti. The inputs were kept as 256*256*32 resolution for the fetal dataset. We have performed an analysis of the performance of the model for both 2D Camus and 3D Fetal datasets. During the experiments, the Adam optimization technique was used and the learning rate was kept at 0.001 with the use of a learning rate scheduler. We have trained the model with 100 epochs. For the training of variational autoencoders, the same hyperparameters were used with 200 epochs. For evaluation and comparison of the results, we have used mean squared error from equation 6 and Dice Score Coefficient from equation 5 were used.

The detailed comparison between the proposed models can be seen in Table 1 and 2. We have also visualized the results for 2D slices which can be seen in the appendix from figure 13, 14, 15, 16. In the figure, the masks were colored according to the overlapping of the pixels where green means true positive, and yellow and red define pixels which are false positive and false negative. The fifth column indicates the overlap of fixed and moving images whereas the sixth column indicates the overlap of the fixed and moved image slices.

6. Discussion

The results are presented for two datasets: 2D CAMUS and 3D Fetal in 1 and 2 respectively. The computation of evaluation metrics between fixed and moving images is referred to as without registration. After registration, evaluation metrics are again computed between fixed and deformed images. The first experiment was done using the baseline model Vanilla-DLIR. We can see that the mean-squared error decreases after registration which indicates that in the case of vanilla-DLIR, the unsupervised registration without considering the anatomy, the similarity between two intensity images increases, but the similarity between fixed and moved masks does not improve satisfactorily or fail in some cases. For that reason, the DSC of the left ventricle and myocardium does not improve much. Figure 13 also shows that the overlapping of the fixed and deformed images is highly irregular containing false positive and negative cases.

Next, we tried to add latent space training to extract the global features using variational autoencoders. In this experiment, we can see the MSE metric decreases as well as the DSC improves than Vanilla-DLIR. Figure 14 also indicates a better overlapping. As VAEs add global context to the learning, model, the results also prove that adding global latent space learning can be beneficial to perform better registration.

The overlapping in the images shows that the boundaries of the regions segmented are irregular or not very smooth. In the third experiment, we tried to add adversarial learning to provide better regularization of the model. From the results both from the table and the images, it can be seen that adversarial learning provides a better regularization and thus also improves the result of vanilla-DLIR.

So, we decided to keep them both in the model and apply them to perform the registration. The results of the proposed AdvAC model outperforms all the previous experiments and thus proved to be the best model working in both the 2D and 3D dataset. Still, there is room for performance improvement. Still, there is room for performance improvement. Hence, we proposed Multi-class Anatomically Constrained and Multi-scale Registration (MACMR) framework which is the best-performing model for the 2D Camus dataset. Although the results on 2D dataset is higher but both 2D and 3D data have the same upward improvement with the proposed models. The fact is that the volume images are low in number for training and also take longer time than 2D for training for each epoch, the result is lower but still satisfactory as this will be the first time temporal registration was done on 3D fetal echocardiography images. In our future plan, we want to add even more 3D data volume to have a better training of the model and also want to apply the multi-resolution framework in case of 3D.

7. Conclusions

The clinical use of echo is still stuck with 2D, likely because doctors can not visualize 3D, but for machine learning it makes more sense to go 3D, for real-time detection with improved accuracy and precision. Existing DLIR or DL echo image processing are all 2D, and so the need for 3D temporal registration for echo images is clearly visible. Also there is less research work done for fetal hearts although the fetal heart can experience congenital heart malformation and functional abnormalities. This thesis focuses on the development of methods for automatic 3D temporal registration for 3D fetal echocardiographic images. The aim was to improve the detection of congenital heart malformations and functional abnormalities in the developing fetus.

One of the two most important aspects of this thesis was to propose a new dataset for fetal echocardiography. 4D volume echocardiography images were collected and annotated with the use of a cardiac motion estimation algorithm. We have conducted several experiments starting with UNET-based DLIR to adding global latent space training with variational autoencoders and adversarial learning to have a better regularization loss.

We have compared the results for both 2D and 3D datasets. The results have shown significant improvements in temporal registration accuracy using evaluation metrics such as Mean Squared Error, and Dice Metric. As the data annotation takes a considerable amount of time, we started the work with a few number of volume images which hindered the overall performance of the 3D dataset. So, we are planning to add more annotated data as well as to evaluate the 3D model in multi-resolution framework.

References

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J.V., Dalca, A.V., 2018. Voxelmorph: A learning framework for deformable medical image registration. IEEE Transactions on Medical Imaging 38, 1788–1800.

Chen, X., Diaz-Pinto, A., Ravikumar, N., Frangi, A.F., 2021. Deep learning in medical image registration. Progress in Biomedical Engineering 3, 012003. URL: https://dx.doi.org/10.1088/2516-1091/abd37c, doi:10.1088/2516-1091/abd37c.

Cheng, X., Zhang, L., Zheng, Y., 2018. Deep similarity learning for multimodal medical images. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 6, 248 – 252.

Ciancarella, P., Ciliberti, P., Santangelo, T.P., Secchi, F., Stagnaro, N., Secinaro, A., 2020. Noninvasive imaging of congenital cardiovascular defects. La radiologia medica 125, 1167 – 1185.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26, 297–302.

Dong, J., Liu, S., Wang, T., 2019. Arvbnet: Real-time detection of anatomical structures in fetal ultrasound cardiac four-chamber planes, in: MLMECH/CVII-STENT@MICCAI.

Fechter, T., Baltas, D., 2019. One-shot learning for deformable medical image registration and periodic motion tracking. IEEE Transactions on Medical Imaging 39, 2506–2517.

Ferrante, E., Dokania, P.K., Silva, R.M., Paragios, N., 2018. Weakly supervised learning of metric aggregations for deformable image registration. IEEE Journal of Biomedical and Health Informatics 23, 1374–1384.

Fu, Y., Lei, Y., Wang, T., Higgins, K.A., Bradley, J.D., Curran, W.J., Liu, T., Yang, X., 2020. Lungregnet: an unsupervised deformable image registration method for 4d-ct lung. Medical physics .

Gong, L., Wang, H., Peng, C., Dai, Y., Ding, M., Sun, Y., Yang, X., Zheng, J., 2017. Non-rigid mr-trus image registration for image-guided prostate biopsy using correlation ratio-based mutual information. BioMedical Engineering OnLine 16.

Green, L., Chan, W.X., Ren, M., Mattar, C.N.Z., Lee, L.C., Yap, C.H., 2023. The dependency of fetal left ventricular biomechanics function on myocardium helix angle configuration. Biomechanics and modeling in mechanobiology 22, 629—643. URL: https://europepmc.org/articles/PMC10097781, doi:10.1007/s10237-022-01669-z.

Haskins, G., Kruecker, J., Kruger, U., Xu, S., Pinto, P.A., Wood, B.J., Yan, P., 2018. Learning deep similarity metric for 3d mr-trus registration. ArXiv abs/1806.04548.

Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T.N., Gleeson, F.V., Brady, M., Schnabel, J.A., 2012. Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration. Medical image analysis 16 7, 1423–35.

Hu, Y., Modat, M., Gibson, E., Ghavami, N., Bonmati, E., Moore, C.M., Emberton, M., Noble, J.A., Barratt, D.C., Vercauteren, T.K.M., 2017. Label-driven weakly-supervised learning for multimodal deformarle image registration. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018) , 1070–1074.

Izzo, R., Steinman, D.A., Manini, S., Faggiano, E., Antiga, L., 2018. The vascular modeling toolkit: A python library for the analysis of tubular structures in medical images. J. Open Source Softw. 3, 745.

Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K., 2015. Spatial transformer networks, in: NIPS.

Jafari, M., Girgis, H., Abdi, A.H., Liao, Z., Pesteie, M., Rohling, R.N., Gin, K., Tsang, T., Abolmaesumi, P., 2019. Semi-supervised learning for cardiac left ventricle segmentation using conditional deep generative models as prior. 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019) , 649–652.

Jafari, M.H., Girgis, H., Liao, Z., Behnami, D., Abdi, A., Vaseli, H., Luong, C., Rohling, R., Gin, K., Tsang, T., Abolmaesumi, P., 2018. A unified framework integrating recurrent fully-convolutional networks and optical flow for segmentation of the left ventricle in echocardiography data, in: Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J.M.R., Bradley, A., Papa, J.P., Belagiannis, V., Nascimento, J.C., Lu, Z., Conjeti, S., Moradi, M., Greenspan, H., Madabhushi, A. (Eds.), Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer International Publishing, Cham. pp. 29–37.

Kingma, D.P., Welling, M., 2014. Auto-encoding variational bayes, in: International Conference on Learning Representations (ICLR).

Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., Lartizien, C., D'hooge, J., Lovstakken, L., Bernard, O., 2019. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE Transactions on Medical Imaging 38, 2198–2210. doi:10.1109/TMI.2019.2900516.

Li, Y., Sun, J., Tang, C.K., Shum, H., 2004. Lazy snapping. ACM SIGGRAPH 2004 Papers .

Lowekamp, B., gabehart, Blezek, D., Marstal, K., Ibanez, L., Chen, D., McCormick, M., Mueller, D., Johnson, H., Cole, D., Yaniv, Z., Posthuma, J., Beare, R., Gelas, A., aghayoor, ltong1130ztr, fsantini, adizhol, Subburam, K., Fillion-Robin, J.C., Anthony, Doria, D., King, B., 2016. kaspermarstal/simpleelastix: v0.10.0. URL: https://doi.org/10.5281/zenodo.168078, doi:10.5281/zenodo.168078.

Mahapatra, D., Antony, B.J., Sedai, S., Garnavi, R., 2018. Deformable medical image registration using generative adversarial networks. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018) , 1449–1453.

Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M.P., Bai, W., Caballero, J., Cook, S.A., de Marvao, A., Dawes, T.J.W., O'Regan, D.P., Kainz, B., Glocker, B., Rueckert, D., 2017. Anatomically constrained neural networks (acnns): Application to cardiac image enhancement and segmentation. IEEE Transactions on Medical Imaging 37, 384–395.

Ong, C.W., Ren, M., Wiputra, H., Mojumder, J., Chan, W.X., Tulzer, A., Tulzer, G., Buist, M.L., Mattar, C.N.Z., Lee, L.C., Yap, C.H., 2020. Biomechanics of human fetal hearts with critical aortic stenosis. Annals of Biomedical Engineering 49, 1364 – 1379.

Painchaud, N., Skandarani, Y., Judge, T., Bernard, O., Lalande, A., Jodoin, P.M., 2019. Cardiac segmentation with strong anatomical guarantees. IEEE Transactions on Medical Imaging 39, 3703–3713.

Rivaz, H., Karimaghaloo, Z., Fonov, V.S., Collins, D.L., 2014. Nonrigid registration of ultrasound and mri using contextual conditioned mutual information. IEEE Transactions on Medical Imaging 33, 708–725.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. ArXiv abs/1505.04597.

Sachdeva, S., Gupta, S., 2020. Imaging modalities in congenital heart disease. The Indian Journal of Pediatrics 87, 385–397.

Simonovsky, M., Gutiérrez-Becker, B., Mateus, D., Navab, N., Komodakis, N., 2016. A deep metric for multimodal registration, in: International Conference on Medical Image Computing and Computer-Assisted Intervention.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image

quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing 13, 600–612.

Wiputra, H., Chan, W.X., Foo, Y.Y., Ho, S., Yap, C.H., 2020. Cardiac motion estimation from medical images: a regularisation framework applied on pairwise image registration displacement fields. Scientific Reports 10.

Wu, G., Kim, M., Wang, Q., Munsell, B.C., Shen, D., 2016. Scalable high-performance image registration framework by unsupervised deep feature representations learning. IEEE Transactions on Biomedical Engineering 63, 1505–1516.

Xie, H., Lei, Y., Fu, Y., Wang, T., Roper, J.R., Bradley, J.D., Patel, P.R., Liu, T., Yang, X., 2022. Inter-fraction deformable image registration using unsupervised deep learning for cbct-guided abdominal radiotherapy. Physics in Medicine and Biology 68.

Yan, P., Xu, S., Rastinehad, A.R., Wood, B.J., 2018. Adversarial image registration with application for mr and trus image fusion, in: MLMI@MICCAI.

Yang, M., Xiao, X., Liu, Z., Sun, L., Guo, W., zhen Cui, L., Sun, D., Zhang, P., Yang, G., 2020. Deep retinanet for dynamic left ventricle detection in multiview echocardiography classification. Sci. Program. 2020, 7025403:1–7025403:6.

Yoon, Y.E., Kim, S., Chang, H.J., 2021. Artificial intelligence and echocardiography. Journal of Cardiovascular Imaging 29, 193 – 204.

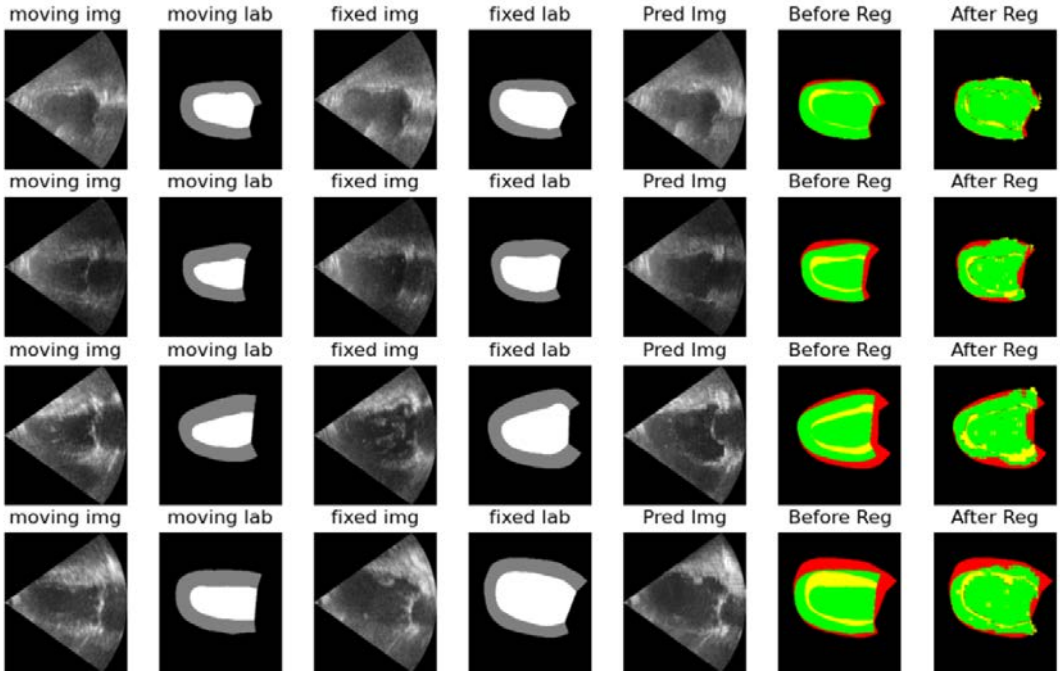


Figure 13: Segmentation Results for Vanilla-DLIR.

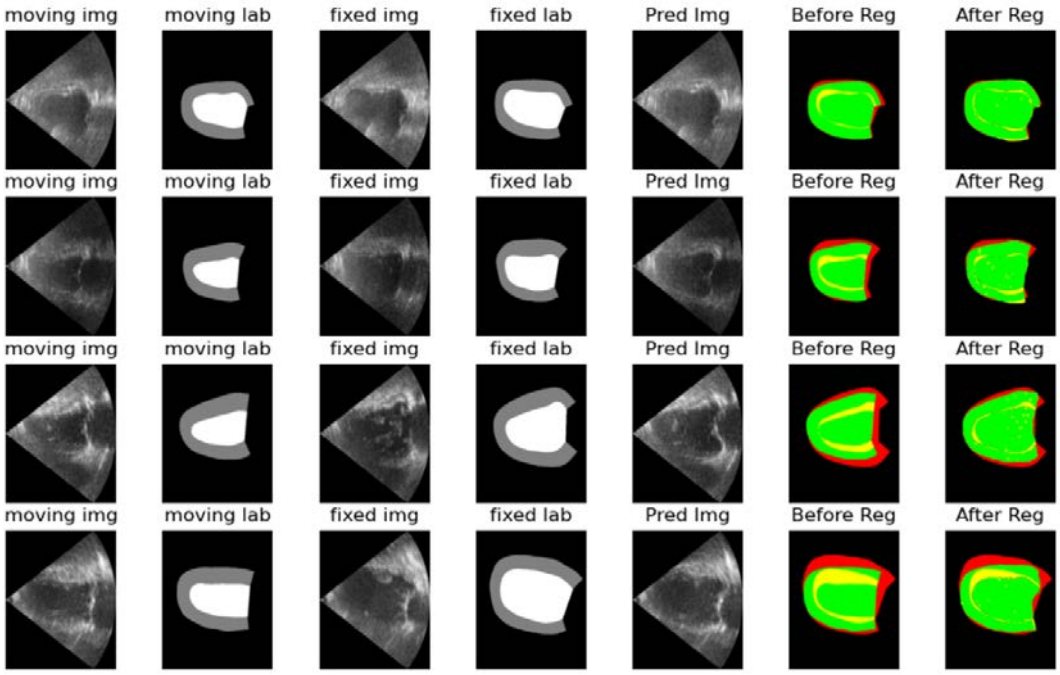


Figure 14: Segmentation Results for AC-DLIR.

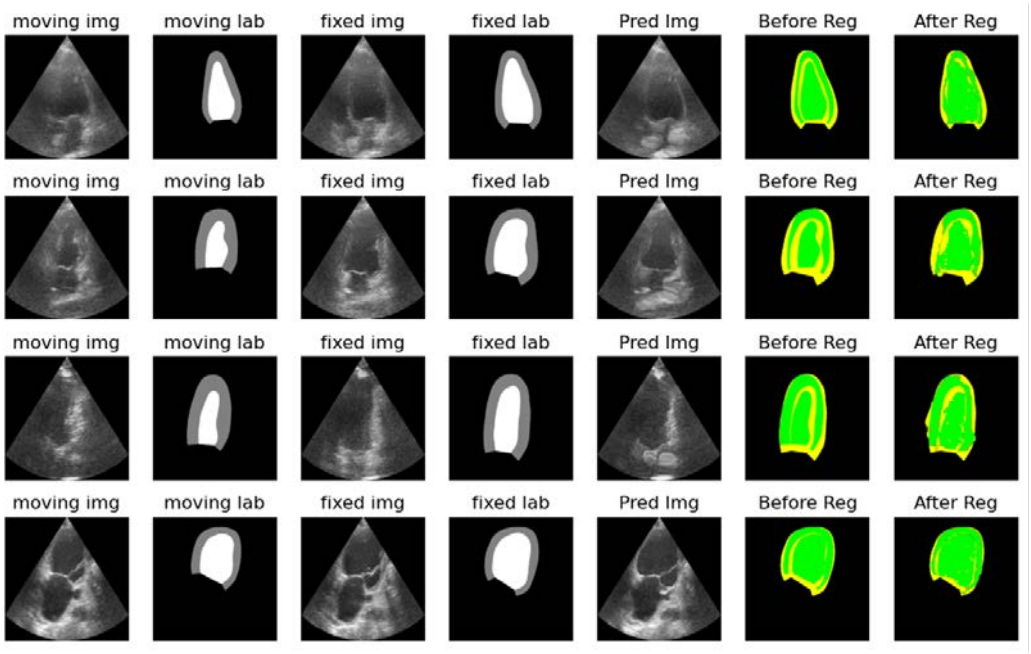


Figure 15: Segmentation Results for Adv-DLIR.

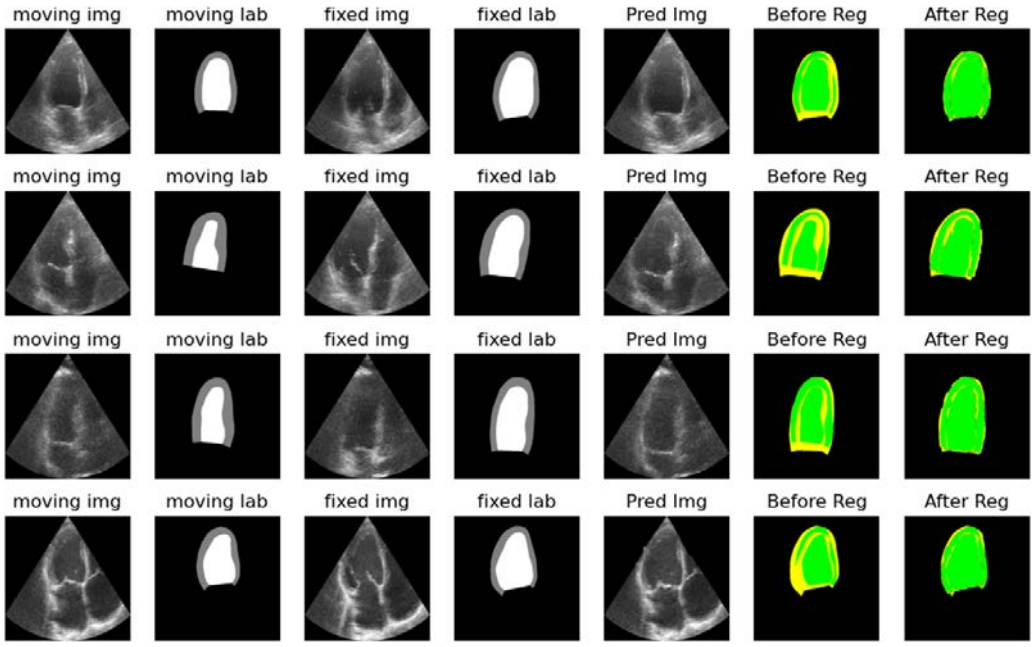
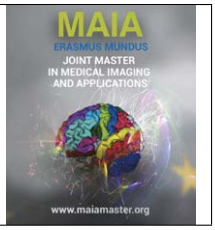


Figure 16: Segmentation Results for AdvAC-DLIR.

Medical Imaging and Applications

Master Thesis, June 2023



X-OOD: How does my model see my data? Robust and Trustworthy Deep Learning for Lung Lesion Detection

Enrique Almar-Munoz[a], Joseph Y. Lo[a]

[a]*Center for Virtual Imaging Trials, Department of Radiology, Duke University School of Medicine, Durham, USA*

Abstract

Introduction - Lung cancer screening using computed tomography (CT) scans plays a crucial role in early detection. This paper addresses challenges in using deep learning models for lung lesion detection, including limited annotated data and algorithm interpretability. Those models tend to be "black boxes" that often fail to generalize. Failures can be caused by many factors, including sub-optimal model development, differences in imaging technologies, or domain shifts in the patient population. As a result of these uncertainties, researchers struggle to improve model performance, and radiologists cannot trust their recommendations. *Methods* - This study analyzes out-of-distribution factors that can influence the model performance in two levels: organ and patient. The impact of OOD factors on the model's performance is assessed by comparing confidence and the rate of considered lesions. That will improve the model's robustness providing explainable insights into how our data is affecting the model. Additionally, we will employ Active Learning techniques, using pseudo-labeling, to enhance the model's generability and performance, further augmenting its capabilities alongside its improved robustness. *Results* - The new included dataset presents similar levels of noise and texture. Regarding the OOD analysis, we spot different data shifts at different levels. OOD distance-based methods present higher accuracy than reconstruction-based ones for our configuration. The uncertainties associated with each OOD factor are mostly similar. The highest difference is found in the OOD organ-level, indicating that those cases confuse our model the most. Active Learning's weak labels improve the model's performance (Average Precision 0.923 with vs 0.8331 without). *Conclusions* - We propose a useful pipeline that ensures good model performance and increases the model robustness by understanding how data shifts can confuse our model. The pipeline is versatile and can be employed to incorporate a new dataset into a study.

Keywords: Lung Cancer, Explainability, Interpretability, Active Learning, Transfer Learning, Out-of-distribution

1. Introduction

Lung cancer, a prevalent form of cancer worldwide, demonstrated the highest percentage of new cases (11.6%) and accounted for the greatest number of deaths (18.4%) among all cancers in 2018 (Bray et al., 2018). Promoting early screening diagnosis is a primary focus in preventing and controlling this disease.

Lesion detection in CT lungs is a critical task that can help to identify and treat lung cancer at an early stage. Deep Learning (DL) has shown great promise in improving the accuracy and efficiency of lung lesion detection (Gu et al., 2021) (Makaju et al., 2018). However, using DL raises concerns about the algorithms' transparency and interpretability (von Eschenbach, 2021).

Detecting tiny pulmonary nodules is a significant challenge. Differentiating them and excluding uncorrelated tissues like bronchi and blood vessels to identify the nodules accurately is hard. Computer-Aided Detection (CAD) system employs a highly sensitive approach to identify nodules. This high sensitivity often results in the formation of candidate nodules with numerous false positives, which presents a major difficulty in the detection process (Setio et al., 2016).

Supervised Deep Learning (DL) requires many labeled training data to make predictions by finding input and corresponding output data patterns. The more labeled training data is available, the more patterns the algorithm can learn and the better it can generalize to new, unseen data.

Medical imaging data is constrained by the limited availability of annotations due to the time-consuming and expensive nature of annotating 3D medical data. Moreover, while medical experts are highly sensitive to the specific condition in question, they are susceptible to inattentional blindness, resulting in elevated miss-rates of unanticipated anomalies and medical conditions.

However, some techniques can be used to reduce the amount of labeled data needed for training, such as transfer learning or active learning (AL). The first involves using a pre-trained model already trained on a large dataset and then fine-tuning it on a smaller, labeled dataset. Nevertheless, it can encounter problems, including dataset bias, domain shift, or model capacity limitations when applied to different data. Active learning proactively selects the subset of examples to be labeled next from the pool of unlabeled data.

Having a model that can effectively adapt to new datasets while maintaining control and comprehending the decision-making process is of utmost importance. Moreover, assessing how potential perturbations in the data may impact the model's performance is crucial. By addressing these factors, we can ensure the development of a robust and reliable model capable of making accurate predictions and facilitating informed decision-making. Our methodology aims to explain the limitations of any model when including a new dataset.

In an ideal scenario, these methods should not rely on domain-specific knowledge or annotated validation sets that are specific to certain cases for optimization. Such reliance could be considered an unwanted form of implicit supervision inherent in the method's design. Therefore, we need to propose a methodology that could be universally applied to different datasets. Concretely, our work will be developed to include Duke's private Data in the multi-center National Lung Study Trial (NLST) (Team, 2011) but could be used for any configuration.

This paper presents a significant contribution, which can be summarized as follows: a pipeline to ensure the robust performance of models when applied to new datasets by identifying the data shifts that can affect the model's confidence. This approach can be applicable to any type of unlabeled data, including signals or videos. The paper's sub-contributions include: (a) the introduction of a methodology that effectively addresses out-of-distribution (OOD) detection in 3D medical images at both organ and patient levels, (b) comparing different datasets in terms of noise and textures to understand the effect on the OOD problem, and (c) the utilization of an active-learning method based on informativeness and representativeness proposing pseudo-label instances to improve model's performance and generalization capability.

2. State of the art

2.1. Out-of-distribution analysis

Modern neural networks have achieved great success, but they are also known to be overconfident even when they encounter inputs with unusual conditions. Finding these inputs is critical to stop models from making uninformed predictions that could endanger neural network applications in the real world. Out-of-distribution (OOD) detection helps to identify differences among data samples, increasing the reliability and safety of a DL model. In an unsupervised manner, its primary objective is to pinpoint unexpected and abnormal data points by learning normal tissue appearance.

Although 2D-OOD is well developed (Berger et al., 2021) (Pacheco et al., 2020), we find few 3D approaches due to increased computational complexity. 3D detection algorithms can be categorized as follows. Density-based methods use an estimation technique to predict probability distribution, distance-based methods measure the proximity among data features, and reconstruction-based techniques calculate the reconstruction error to spot data dissimilarities. To our knowledge, purely density-based methods have not been explicitly utilized in medical imaging, presumably because they do not provide an anomaly score at the pixel level. We, therefore, focus on reconstruction- and distance-based methods below.

2.1.1. Reconstruction-based methods

Several methods have been proposed to address the issue at hand. (Shyu et al., 2003) presented a novel approach utilizing Principal Component Analysis (PCA)-based reconstruction. Using DL, (Schlegl et al., 2017) employed an iterative back-propagation method within a Generative Adversarial Network (GAN) framework. It is worth noting that autoencoders (AE) reconstruction methods offer notable advantages in handling non-linear data relationships and enabling pixel-wise detection.

A study conducted by (Chen and Konukoglu, 2018) demonstrated the effectiveness of combining a Variational Auto-Encoder (VAE) with an adversarial loss applied to the latent variables. This approach resulted in improved performance by leveraging a pixel-wise reconstruction error. Building upon this notion of error, (Zimmerer et al., 2019) utilized various Auto-Encoders (AEs) specifically designed for brain computed tomography (CT) scans. In related work, (Alain and Bengio, 2014) provided evidence that AEs have a tendency to learn a condensed representation of the underlying data distribution by capturing the derivative of the log-density with respect to the input.

Nevertheless, using only the reconstruction error for scoring overlooks the reconstruction model's internal representation and lacks formal claims and comparability between samples. To address this, (Zimmerer

et al., 2018) combined reconstruction with density-based scoring on the Context-encoding Variational Autoencoder (ceVAE). It utilizes a context-encoding mechanism to encode contextual information of input data and a VAE to learn the underlying data distribution.

2.1.2. Distance-based methods

Various methods have been employed to evaluate the similarity between a test instance and the distribution of training instances. Among them, the most commonly employed scoring metric is the Mahalanobis distance. This distance measure considers the covariance structure of the training instances, providing a comprehensive evaluation. In a study by (Karimi and Gholipour, 2022), singular value decomposition (SVD) was employed on the network features. By extracting singular values, an image embedding was generated. The OOD score was then determined as the distance between a test sample and its nearest neighbor in the training set.

Other groups, such as (Hendrycks and Gimpel, 2016), pointed to using the maximum softmax probability (MSP) for the detection. ODIN (Liang et al., 2017) achieved further improvement over MSP by incorporating temperature scaling of softmax outputs and input perturbations. However, a major limitation of ODIN is that it requires the availability of samples to select the temperature scaling factor and the magnitude of the input perturbations. This issue was addressed by (Hsu et al., 2020) by proposing a generalized, G-ODIN, eliminating the fine-tuning necessity.

2.2. Active Learning

By explaining the impact of new data on the model, we can enhance its resilience. Furthermore, incorporating new data into the training process not only enhances the model's generalizability but also boosts its overall effectiveness.

Active learning endeavors to streamline the data collection by automatically discerning the instances that necessitate expert annotation for efficient and effective model training. Its objective is to minimize labeling effort while maximizing the performance achieved by the machine learning algorithm.

It has demonstrated success in various domains, including image classification (Beluch et al., 2018) (Sener and Savarese, 2017), object detection (Bengar et al., 2019), regression (Käding et al., 2018), and semantic segmentation (Golestaneh and Kitani, 2020) (Wang et al., 2020).

AL strategies can be categorized into three main groups: informativeness (Bengar et al., 2021) (Cai et al., 2014) (Gal et al., 2017) (Guo, 2010) (Yang et al., 2015), representativeness (Saito et al., 2015) (Sener and Savarese, 2017), and hybrid approaches (Yang and Loog, 2018) (Huang et al., 2010). The informativeness criterion selects samples that exhibit high uncertainty, thereby impacting the model's generalization capability (confusing the classifier). Representativeness ensures the inclusion of diverse samples that align with the underlying data distribution.

Although active learning has been extensively investigated in the field of classification tasks, it has garnered comparatively less attention in the domain of deep object detection (Brust et al., 2018). With that purpose, (Kao et al., 2019) introduced a ranking approach for images based on the localization tightness and stability criteria. Localization tightness measures the compactness of detected bounding boxes, while stability estimates their robustness in both the original image and a noisy version of it. Additionally, (Brust et al., 2018) employed the computation of marginal scores (Ronald J. Brachman) for candidate bounding boxes and incorporated them using various merging functions.

We propose a method based on informativeness and representativeness combined with pseudo-labels. Pseudo-labeling based on the AL results is not novel. For instance, in the context of image classification, the definition of pseudo-labels varied across studies. In (Lee et al., 2013), the pseudo-label was defined as the class with the highest probability. Conversely, (Bank et al., 2018) introduced multiple techniques to derive the confidence measure for pseudo-labels. (Zotova et al., 2019) showed that pseudo-labeling gives small further improvements for a segmentation task. But no one proposed an AL algorithm using pseudo-labels for detection.

2.3. Explainable AI

The incorporation of explainability into deep learning models serves as a means to tackle various obstacles, including the issues arising from data shift. Data shifts can result in diminished performance and untrustworthy predictions. Through the comprehension of the decision-making process facilitated by explainability, data scientists can pinpoint the specific features, patterns, or data attributes that the model depends

Explainable AI, XAI, focuses on creating artificial intelligence (AI) systems that provide transparent and interpretable explanations for their decisions. Its goal is to enhance AI models' understanding, trust, and accountability by providing insights into the underlying factors, logic, and reasoning behind their outputs.

Most of the existing research in the field of explainable AI (XAI) has heavily relied on using Saliency Maps as a common method for providing explanations. However, emerging studies have demonstrated that they may lack stability, meaning they can vary significantly in their output and may not consistently highlight the most relevant features or areas of importance in the input data. This instability raises concerns about their reliability and robustness of them as the sole method for explainability.

(Arun et al., 2021) analyzed the performance of eight commonly used saliency map techniques regarding (a) localization utility (segmentation and detection), (b) sensitivity to model weight randomization, (c) repeatability, and (d) reproducibility. They proved that all eight saliency map techniques failed at least one of the criteria and were inferior in performance compared with localization networks. Figure 2 shows the results of each saliency method for the detection task and the predicted output of RetinaNet (RNET).

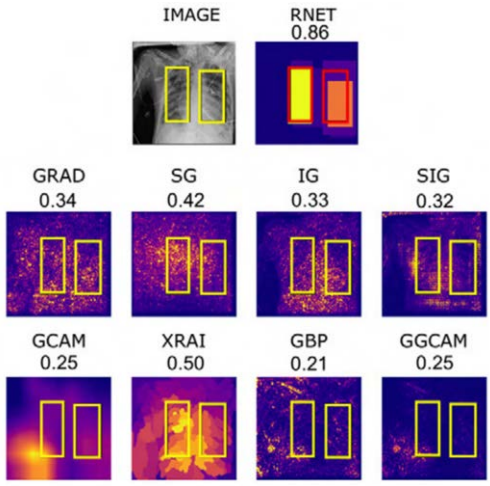


Figure 1: Example saliency maps for RSNA pneumonia dataset with corresponding utility scores. GBP = guided backpropagation, GCAM = gradient-weighted class activation mapping, GGCAM = guided GCAM, GRAD = gradient explanation, IG = integrated gradients, SG = Smoothgrad, SIG = smooth IG (Arun et al., 2021)

Consequently, researchers are actively exploring alternative approaches and techniques within XAI to address these limitations. We hypothesize that OOD detection could be used for the model's explainability by providing insights into how a deep learning model makes predictions and how data shifts affect the model's performance. We can better understand the model's decision-making process and assess its reliability.

For example, if the model is presented with input outside the distribution it was trained on and provides a high-confidence prediction, this could indicate that the model is overconfident and may be making unreliable predictions. Conversely, if the model identifies OOD inputs and produces lower confidence predictions, this could be a sign that the model is aware of its limitations and is cautious in its predictions.

Utilizing the principle of explainability, our approach aims to establish a heightened sense of trustworthiness in the context of federated learning relating OOD factors with each associated performance. In a clinical study with multi-center data, trustworthiness assumes utmost significance.

2.4. Explainable OOD Analysis

By understanding the reasons behind OOD data, researchers can identify data issues, improve models, build trust, and promote equitable decision-making. Explaining OOD data is essential for reliable predictions and accurate assessments in various fields.

(Hendrycks and Gimpel, 2016) utilizes confidence scores, maximum predicted probability, from the softmax layer to identify misclassified OOD cases. Recently, (Xu-Darme et al., 2023) proposed method emphasizes interpretability, aiming to provide insights into why certain samples are identified as OOD. The authors introduce an auxiliary network that learns to identify the patterns contributing to the OOD detection decision. Other authors used OOD cases to interpret the model using saliency maps (Fong and Vedaldi, 2017).

But all of them have focused on understanding how the OOD cases are classified. We propose an explainable OOD analysis aimed at identifying which OOD cases exert the most negative influence on our model.

3. Material

3.1. Dataset

- **LIDC/IDRI dataset**: It is a publicly available thoracic standard and low-dose computed tomography (CT) dataset. It consists of 601 labeled cases, each including a set of CT images and annotations of lung nodules by four experienced radiologists. They include information on nodules' location, size and shape. The dataset also includes assessments of the probability of malignancy and radiologists' confidence level in nodule detection.

- **Duke private dataset**: It includes 7345 lung low-dose CT, but only 1.55% are labelled. They included a radiologist inform with patient information. They have been resampled to match LIDC's shape.

3.2. LUNA16 Model

In our study, we have utilised the MONAI model for the LUNA16 Challenge, specifically the model that achieved the second position in the challenge ranking. The chosen model has been made publicly available and is built upon the RetinaNet network architecture. Utilizing the MONAI model aligns with our research objectives and allows us to benefit from the model's strengths and advancements.

RetinaNet is a prominent object detection framework widely employed for accurately detecting objects in images. It tackles the challenge of detecting objects at multiple scales by introducing a novel component termed Focal Loss, which effectively prioritizes the training of challenging samples. In Figure 2, we can see RetinaNet network architecture.

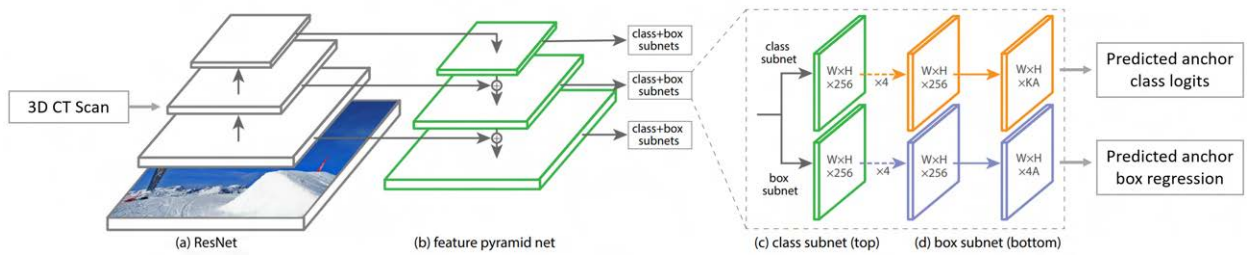


Figure 2: RetinaNet utilizes a ResNet backbone (a) in conjunction with a Feature Pyramid Network (FPN) (b) (Lin et al., 2017). Two subnetworks are attached to this backbone: one for classifying anchor boxes (c) and another for regressing from anchor boxes to ground-truth object boxes (d). (Lin et al., 2018)

The fundamental concept underlying RetinaNet revolves around utilizing a single deep neural network to concurrently predict object bounding boxes and classify their corresponding object categories. This enables efficient and accurate object detection.

At its core, RetinaNet builds upon a feature pyramid network (FPN) architecture (Lin et al., 2017). FPN capitalizes on the multi-scale features extracted from various levels of a backbone network, ResNet, to achieve robust object detection across different object scales. (He et al., 2016)

RetinaNet presents a novel mechanism called anchors, which generate a set of fixed-size bounding boxes at each spatial location in the feature map. These anchors serve as reference points for the network to predict object bounding boxes. It enables the detection of objects with varying sizes and shapes by employing anchors with diverse scales and aspect ratios. (Lin et al., 2018)

To predict object presence and its corresponding class, RetinaNet employs two parallel sub-networks: the classification subnet and the regression subnet. The classification subnet estimates the probability of an anchor containing an object of a specific class, while the regression subnet computes refined bounding box coordinates for each anchor.

One of the primary challenges in object detection is the significant imbalance between background and foreground samples. Most anchors do not encapsulate any objects of interest, resulting in many easily classified negative samples during training. This imbalance causes the network to be biased toward background predictions, leading to suboptimal performance.

To address this issue, RetinaNet introduces the concept of focal loss, which mitigates the contribution of easily classified samples and emphasizes challenging ones. It accomplishes this by assigning higher weights to misclassified examples and reducing the weight for well-classified ones. This mechanism enables the network to prioritize learning difficult samples, which is crucial for achieving accurate object detection.

Focal loss (Lin et al., 2018) is formulated as a modification of the standard cross-entropy loss function. It introduces a tunable parameter known as the focusing parameter, which governs the degree of emphasis placed on hard examples. By appropriately adjusting this parameter, the loss function can be tailored to balance false positives and false negatives, depending on the application's specific requirements.

RetinaNet minimizes the combined loss from the classification and regression subnets during the training phase. This joint optimization enables the network to simultaneously learn accurate object detection and precise object localization through bounding boxes.

In our implementation, we perform inference on patches if the input image exceeds GPU memory capacity. Data-loader retrieves boxes, and data augmentation is applied to these boxes. We use a batch size of 1, shuffle the data, and employ 7 workers. The model is trained for 300 epochs to facilitate convergence and optimal performance. We use 0.01 as the learning rate until epoch 160, which is reduced to 0.001.

4. Methods

The proposed pipeline is shown in Figure 3.

We will conduct an out-of-distribution study and dataset profiling using the Duke Dataset to identify potential factors influencing our model's performance.

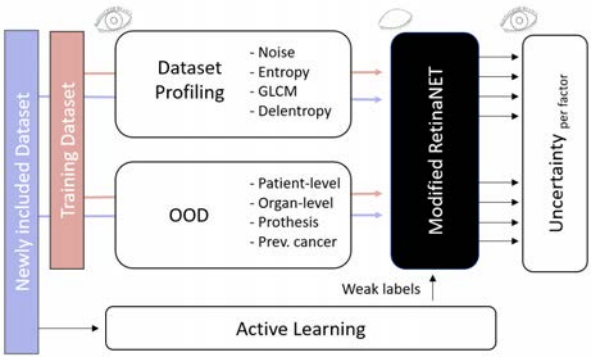


Figure 3: Proposed pipeline to improve model's robustness. As the DL model is a black box, we are going to spot, on the left side of it, potential data factors that may confuse the model. We are going to follow up on those cases through the model, and on the other side of it, we will interpret how the model sees our data.

For the out-of-distribution study, we will utilize the LIDC-IDR data as the reference distribution, which was used to train the model. Our study tries to identify Duke cases demonstrating deviations from the training distribution in two levels. In parallel, we will utilize an Active Learning approach to propose a subset of Duke cases to be labeled, maximizing the model's performance. Previously, transfer learning was applied to predict Duke cases with the model trained on LIDC cases. Moreover, the AL algorithm will provide us with certainty-case-level scores. The scores and the cases spotted in each OOD factor are used to find the factor-level certainty scores. The intersection of these two lines of investigation represents a crucial point.

This analysis allows us to understand the impact of different data abnormalities on the model's performance, increasing the robustness. Additionally, AL-labeled and pseudo-labeled cases improved model generalization capabilities because the RetinaNet model will be trained with more data. Afterward, we will proceed to elucidate the pipeline methodology in stages.

4.1. Data Study

This section will analyze the data independent of its ground truth or predicted labels through two distinct studies. While the first study (section 4.1.1) focuses on studying each dataset individually in terms of texture and noise levels. The second one (section 4.1.2) studies Duke's distribution in relation to LIDC's distribution.

4.1.1. Dataset Profiling

Noise estimation. Ideally, training and test data should have the same noise level. If the level of noise in the test data is different, the model may not be able to generalize well and may fail to classify or predict new data points accurately. A common definition of image noise is the standard deviation (SD) of the measured Hounsfield units (HU) in a physically homogeneous volume (Bongartz, 1999). In chest CT, optimal representation of image noise may be obtained by segmenting the entire tracheo-bronchial tree lumen and measuring the SD of this air (Wisselink et al., 2021). As the trachea is a homogeneous volume, if the deviation is low, there is less presence of random fluctuations or variations in the signal due to noise.

We used TotalSegmentator (Wasserthal et al., 2022) to segment the trachea volumes from both datasets. This nnU-Net model was pretrained in 1204 CT scans and segments 104 structures. Once we obtained the masked volumes, we operated the coefficient of variation of the trachea volume (Equation 1). In order to mitigate the influence of the trachea contour on the noise measure, we applied a 3D-erosion operation to the trachea volumes using a disk with a radius of 5 pixels.

$$CV(\%) = \frac{\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}}{\bar{x}} \times 100 \tag{1}$$

Texture analysis. When comparing image datasets, the common practice is to evaluate metrics like the total number of images, the number of images in each class, or class distributions in the dataset. However, all these metrics are defined by humans and do not provide insights about the underlying data distribution. We want to analyze whether this aspect of complexity impacts the ability of a neural network to learn from that dataset. With that purpose in mind, we will use a similar approach to (Rahane and Subramanian, 2020) to analyze the texture complexity. They used quantitative metrics to identify which dataset is more complex or harder to "learn" concerning a deep-learning-based network. They studied four video datasets from the autonomous driving research community; we will adapt the algorithm to CT volumes. Definition of the used-study metrics:

- Shannon Entropy: The higher the entropy value is, the more information is required to describe or transmit it. It entirely relies on individual pixel values

  $$H(X) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{2}$$

  In this equation, X is a discrete random variable with n possible values, and p_i is the probability that X takes on the value x_i.

- GLCM: Statistical method used to describe the spatial relationship between pairs of pixels in an image.

  $$GLCM = -\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} p(i,j) \log p(i-j) \tag{3}$$

  where n is the number of gray levels, and $p(i-j)$ is the probability of two pixels having intensities i and j, separated by the specifies offset.

- Delentropy: It is based on the probability density function deldensity (Larkin, 2016). This density distribution uses spatial image and pixel co-occurrence. The usage of gradient vectors in this entropy allows for global image features to be considered and capture non-local information.

- UMAP: We use this dimensionality reduction technique to represent the GLCM matrix visually. The first phase consists of constructing a fuzzy topological representation by using simplices. Then, we optimize our embedding (by using stochastic gradient descent) to have as close a fuzzy topological representation as possible (measured by cross entropy). We use additional parameters like the number of nearest neighbors in UMAP to show how local and global structures change/shift differently in different datasets.

4.1.2. *Out-of-distribution detection*

To address the ambiguity surrounding "data outside the distribution", which relies on the definition of a "normal" patient, we conducted a comprehensive analysis at both the patient-level and organ-level. At the patient-level, we examined the entire CT scan to identify potential shifts caused by protocol variations and noise/artifacts. At the organ-level, our focus was on prioritizing anatomical differences. We expected to find different distribution shifts at both stages. But, we hypothesized that the classified organ-level OOD cases would affect the model's performance more. Nevertheless, patient-level ones can also affect it because the model's input is the unmasked CT volumes.

4.1.2.1. *Patient-level-out-of-distribution.*

As detailed in section 2.1, we discussed different categories of out-of-distribution (OOD) methods. We proposed a distance-based and a reconstruction-based method to determine the superior option for our configuration.

Distance-based algorithm: Histogram features As in many medical imaging tasks, our training data presents semantic homogeneity as it consists of chest CT scans. As a result, the intensity histograms of two images with different semantic meanings can be easily differentiated. Since the primary sources of out-of-distribution (OOD) data stem from semantic and covariate shifts, histograms provide a robust alternative to deep learning (DL) methods in this particular task (Frolova et al., 2022). Semantic data shift refers to a change in the meaning or interpretation of the data, such as CT scans of other body parts. Covariate shifts refer to a change in the distribution of the input features, including variations in imaging equipment, imaging protocols, patient demographics, or pathologies. Moreover, as our study task of nodule detection represents a focal disease that occupies only a very small portion of the volume, patients with the pathology would not be categorized as OOD. However, if they exhibit more diffuse pathologies like pneumonia, they may be classified as OOD.

We proposed a straightforward approach based on image histogram features comprising two steps. Firstly, we compute the histogram features for the LIDC dataset and identify its distribution center. Secondly, we calculate the histogram features for the Duke Dataset and utilize the Mahalanobis distance to measure the deviation from the center of the LIDC distribution (Figure 7). Mahalanobis distance is defined as:

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

Where x is the test instance, μ is the mean of the training instances, Σ is the covariance matrix of the training instances, and $D_M(x)$ is the Mahalanobis distance between x and the training distribution.

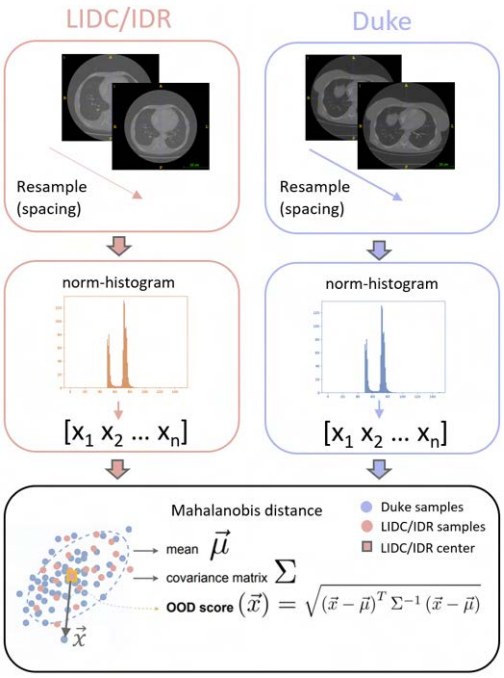


Figure 4: Starting from both datasets, we resample them and get the histograms normalized in intensities and size. After we find the Duke's instance distances to the LIDC distribution center.

The primary objective of this methodology is to identify distribution outliers without explicitly categorizing their underlying causes. However, by determining which region of bins is the most disparate between the center training data histogram and validation cases' histograms, we can cluster certain reasons. That can be done using weighting techniques such as multiplying ramps or exponential curves to the histograms. As an example, we have proposed a modification specifically targeted at detecting the presence of prostheses and artifacts within the Duke Dataset, as illustrated in Figure 5. For that, we first subtracted the center LIDC histogram from each case in the Duke Dataset. If a CT scan contains a metallic prosthesis or a white artifact, there will be pixels at the latest bins of the histogram. To emphasize the contribution of the brightest pixels, we multiply by an exponential function with an empirically chosen function, in our case $e^{8(x-1)}$, to assign greater importance to the brightest region. This way, the algorithm classifies based on the prosthesis's area and brightness.

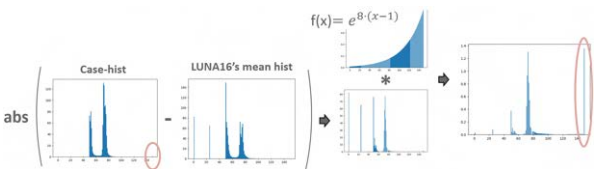


Figure 5: Pipeline for the prostheses/artifacts detection algorithm.

Reconstruction-based algorithm: Variational Auto-Encoder

A Variational Auto-Encoder (VAE) is a generative model that combines ideas from auto-encoders and variational inference. It learns a latent representation of the input data by jointly training an encoder and a decoder neural network. The encoder network maps an input data point to a probability distribution in the latent space. This distribution is typically assumed to follow a multivariate Gaussian distribution, with the mean and variance parameters predicted by the encoder network. The decoder network takes a sample from the latent space and reconstructs the input data point. The objective of the VAE is to maximize the reconstruction accuracy while also encouraging the learned latent space to follow a desired prior distribution, often a standard Gaussian distribution.

In our configuration, the VAE employs a combination of a reconstruction loss and a regularization term known as the Kullback-Leibler (KL) divergence during training. The KL divergence measures the difference between the predicted latent distribution and the desired prior distribution. By jointly optimizing the reconstruction loss and the KL divergence, the VAE learns to encode the input data into a meaningful latent space and generate reconstructions that closely resemble the original input.

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] + \mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right] \tag{4}$$

In this equation, $\mathcal{L}$ represents the objective function of the VAE, θ and ϕ are the model parameters, $\mathbf{x}$ is the input data, $\mathbf{z}$ is the latent variable, $q_\phi(\mathbf{z}|\mathbf{x})$ is the approximate posterior, $p_\theta(\mathbf{x}|\mathbf{z})$ is the likelihood, and $p(\mathbf{z})$ is the prior. The term $\mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right]$ represents the Kullback-Leibler divergence between the approximate posterior and the prior. The model's architecture can be found in figure 6.

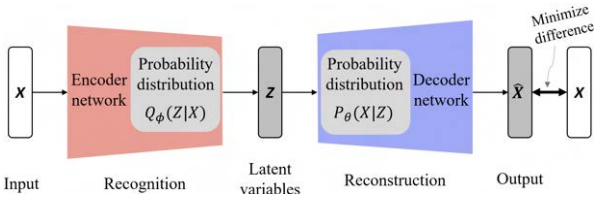


Figure 6: VAE generation model consisting of an encoder network $Q_\phi(Z|X)$ and a decoder network $P_\theta(X|Z)$.

We trained the VAE using the LUNA16 cases as our training dataset. Once the VAE had been trained, we utilized it to reconstruct Duke cases. The reconstruction error was then calculated by comparing the reconstructed Duke cases to their original counterparts. That error serves as the out-of-distribution score, indicating how well the VAE can reconstruct the Duke cases compared to the training data. A higher reconstruction error suggests a larger deviation from the training distribution, implying a higher likelihood of the Duke cases being out-of-distribution, unseen, or anomalous. The training process will be carried out with a batch size of 6, utilizing a learning rate of 0.01 for a total of 200 epochs. Additionally, weight decay regularization will be applied with a value of 5e-7.

4.1.2.2. Organ-level-out-of-distribution.

To specifically analyze data shifts in the studied anatomy, we performed lung volume segmentation using the TotalSegmentator model (Wasserthal et al., 2022). This model segments lung volumes into five distinct structures: the two superior lung lobes, the two inferior lobes, and the middle right lobe.

We merged the 5 lung lobes and applied the masks to the original CT volumes. By reducing the influence of extraneous factors, we enable a more precise examination of the underlying anatomical characteristics and reduce medical imaging acquisition shifts.

In addition to utilizing histograms to capture intensity-based information, we introduced a novel approach to incorporate anatomical information through histograms of oriented gradients (HOG), computed by:

$$V_{mag-grad.} = \sqrt{\mathrm{SF}(V,x)^2 + \mathrm{SF}(V,y)^2 + \mathrm{SF}(V,z)^2} \tag{5}$$

$$HOG_{features} = histogram(V_{mag-grad.}, n_{bins}) \tag{6}$$

Let SF represent the Sobel filter applied to a 3D CT volume to extract information in a specific direction.

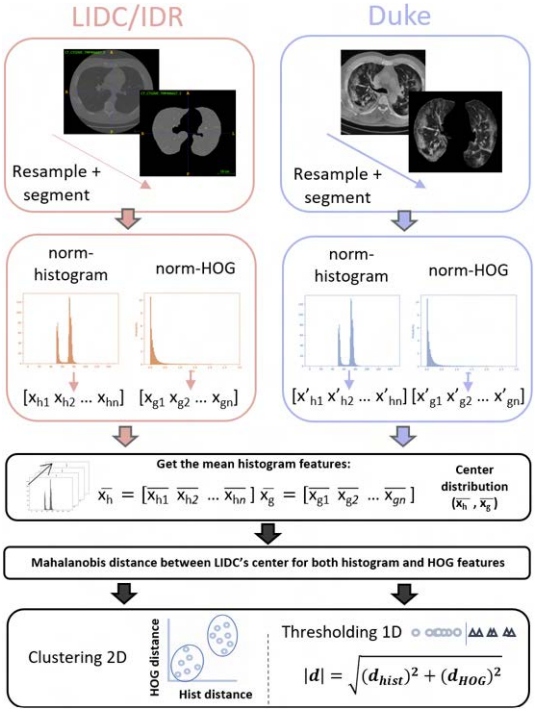


Figure 7: Pipeline for organ-level out-of-distribution analysis incorporating histogram and HOG features. The bottom box recaps the distinction between qualitative and quantitative methods.

HOG integration also provides valuable insights into the signal's energy. However, using HOG features for the complete CT volume doesn't improve the classification because it is very sensitive to variations in the patient's body morphology. We use the Mahalanobis metric between LIDC's center and Duke's instances for both histogram features and HOG features. The proposed pipeline is shown in Figure 7

The patient-level analysis differs primarily in the utilization of 2D information. In this regard, we proposed a qualitative analysis employing density-based spatial clustering of applications with noise (DBSCAN), which utilizes distance measurement, typically Euclidean distance, and a minimum number of points to group data points. A notable feature of this algorithm is its ability to identify outliers, as it effectively captures points residing in low-density regions. We will identify the instances of those regions as OOD cases. Additionally, we also introduced a quantitative method that involved calculating the Euclidean distance between the center of the LIDC dataset and the Duke samples.

4.2. Active Learning

Figure 8 shows the implemented AL strategies.

Baseline: Random Sampling: From the 7231 unlabeled Duke scans, 16 cases are randomly selected and added to the pipeline during each iteration.

AL Strategy 1: Uncertainty Sampling: We utilize the model trained on LIDC cases to make predictions for the 16 unlabelled cases in each iteration. The model provides certainty scores at the lesion level. However, in order to rank the scans based on certainty, we need to calculate a patient-level score. The most challenging aspect is transitioning from lesion-level to patient-level uncertainties. This is achieved by computing the mean of all lesion certainties per scan prior to filtering them to exclude potential false-positive detections. The filtering criteria include (1) size, where lesions with no axis exceeding 3 mm (as determined by radiologists) are discarded; (2) certainty score, with a threshold of 0.1; and (3) mask, where lesions whose predicted center falls outside the lung volume are considered. A cautious approach is taken to prevent false-negative lesions near the lung boundaries or pleura by dilating the lung masks using a 15-pixel radius disk element. At every iteration, the two scans with the highest confidence score are added to the "certain" label pool, while the two scans with the weakest certainty are added to the "uncertain" pool.

AL Strategy 2: Representativeness: The manual labeling process is also leveraged by introducing a recency condition. Only the most recent cases per patient in the uncertain cases pool are added to the annotated pool. To identify the most recent scans per patient, we thoroughly examined patients' information.

AL Strategy 3: Pseudo-Labelling: To define the pseudo-labels, from the certain pool, we re-filter the lesions with more stringent criteria. Each lesion must be inside the dilated lung lesion, have all axes greater than 3 mm, and have an associated certainty score higher than 0.3. Additionally, non-maximum suppression (NMS) is applied to reduce overlaps in bounding boxes and eliminate redundancy in object detection. Those labels will be used to retrain the model.

5. Results

5.1. Dataset characterization

Similar entropy values were observed among the datasets in the **texture analyses**. Table 1 presents the average values for each dataset and corresponding entropy type. GLCM reveals a prominent distinction, with Duke Data standing out by showcasing a higher value.

Dataset	Shannon	Delentropy	GLCM
LUNA16	**8.505 (0.205)**	**3.550 (0.627)**	2.967 (1.151)
Duke	8.501 (0.110)	3.412 (0.736)	**3.034 (1.300)**

Table 1: Analysis of entropy measures across datasets and metrics

We employed GLCM and UMAP visualization plots to provide a more intuitive image space representation (Figure 9). "Difficult" datasets are more densely distributed. By utilizing this feature space, we reduced dimensionality based on the topology of texture-based features. Varying numbers of nearest neighbors were utilized to capture different complexity levels. Our findings align with the metrics presented in Table 1, indi-

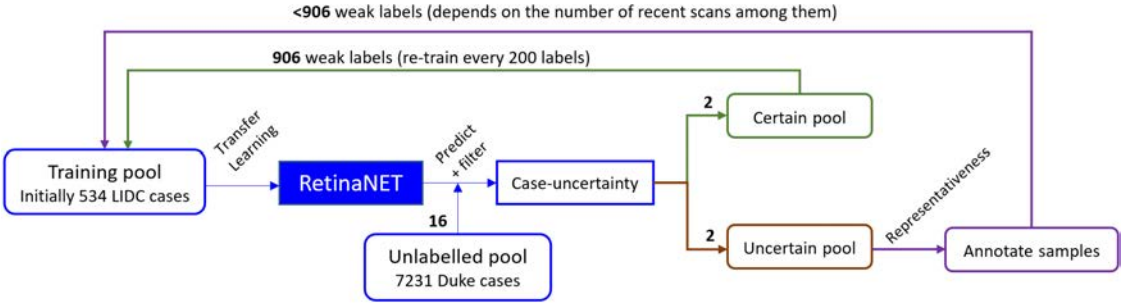
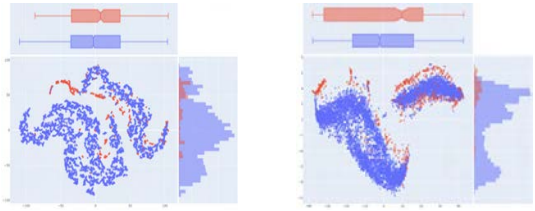


Figure 8: Active learning setup showing uncertainty sampling, representativeness filtering, and automated pseudo-labeling.

cating that the Duke Dataset demonstrates greater complexity in higher orders (global information).



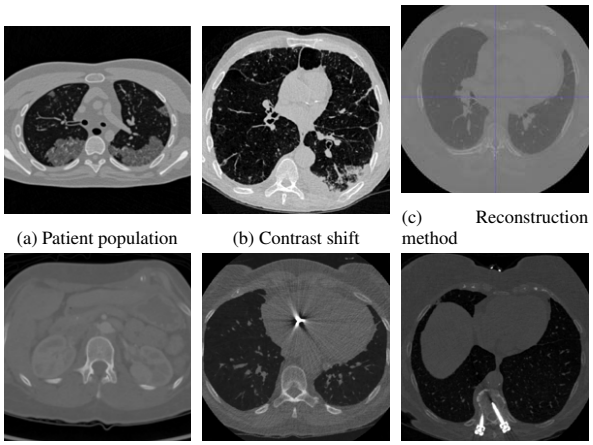
(a) Neighbors: 20 (b) Neighbors: 500

Figure 9: Two-dimensional projection of each dataset using UMAP. The plots show embedding with the distribution of each dimension on each axis. Duke Dataset (blue) and LUNA16 (red).

Noise estimation was conducted by applying the coefficient variation to the segmented trachea volumes (Figure 12 highlights the regions represented by the color red). Duke Dataset exhibits a coefficient variation of **43.37** ± **14.08**, while the LUNA16 dataset shows a value of 42.16 ± 13.39. Although both datasets demonstrate comparable noise levels, Duke data exhibits a slightly higher level.

5.2. Out-of-distribution analysis

The graphical representation, denoted as Figure 10, illustrates the distinct classifications of shifts among the data samples derived from the Duke dataset. These shifts have been identified using histogram features, specifically employed on the entire CT volume.



(a) Patient population (b) Contrast shift (c) Reconstruction method

(d) Another body-part (e) Artifact presence (f) Prosthesis presence

Figure 10: Classification of shifts in Duke Data Samples using Histogram Features on total CT volumes.

For our prosthesis/artifact detection approach using weighted histogram features, out of the total 7345 scans, 338 scans (4.6% of the dataset) are identified as belonging to the prosthesis group. To assess the accuracy of our approach, we conducted a thorough manual verification process on 50 randomly selected cases from the prosthesis group. It confirms the presence of a prosthesis or artifact in 47 instances, providing strong evidence for our method's efficacy. Most of the prostheses detected are Posterior Spinal Fusion (PSF), but we also found reverse shoulder prostheses and cardiac prosthetic valves.

Regarding the reconstruction-based method, VAE's training is stopped once the validation loss does not decrease for more than 3 epochs. Fig 11 presents a selection of reconstructed volumes during this training process.

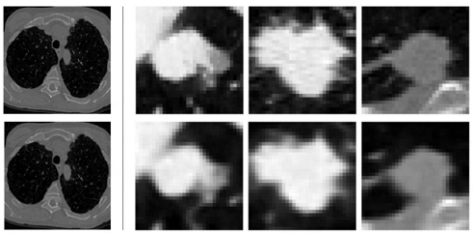


Figure 11: Original CT (top row) and VAE reconstructed volumes (bottom row). (a) Shows the entire volume, (b-d) zoom in on nodules.

Due to the unavailability of OOD ground truth (each scan has a label in/out distribution), we adopt a methodology to determine the correctness of their classification wherein we select the 100 instances farthest away from each method's training distribution and manually examine if they are OOD. Our evaluation process entails categorizing these instances into three distinct groups. The first group is labeled as *outside training distribution* and comprises instances with explicit reasons for not belonging to the training distribution. These instances exhibit significant dissimilarities and discernible patterns that distinguish them from the training data. The second group, termed *partially outside training distribution*, consists of cases that exhibit certain peculiarities, yet there exists a plausible explanation as to why our algorithm classified them as non-conforming to the training distribution. Although these instances deviate somewhat, their dissimilarities are insufficient to categorize them as distinctly different. The third group, denoted as *inside training distribution*, encompasses instances that demonstrate no apparent reasons for being classified as outside the training distribution. These instances exhibit consistency with the patterns and characteristics observed within the training data. Results are shown in Table 2. We can observe that the proposed distance-based algorithm works better for our configuration.

Labels	HIST	VAE
Outside training distribution	**0.85**	0.71
Partially outside training distribution	0.08	**0.17**
Inside training distribution	0.07	**0.12**

Table 2: Comparative Classification Accuracy of Out-of-Distribution Patient-Level Methods: distance-based (Histogram Features, HIST) vs. reconstruction-based (Variational Autoencoders, VAE).

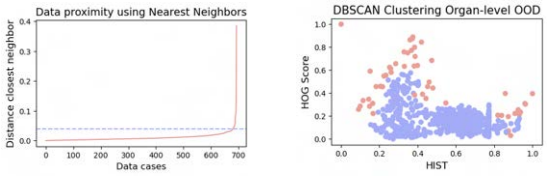
Descending on the organ level, Figure 12 shows the segmented volumes (5 lung lobes and trachea).

Figure 12: TotalSegmentator's segmentation in coronal, axial, and 3D view. Both superior lobes, both inferior lobes, and the right middle lobe are represented in blue scale. The trachea structure shown in red.

Before applying the DBSCAN model, it is necessary to determine two parameters:

- Minimum number of points needed to consider a new cluster: 6. It is 2 times the data dimension.
- Epsilon (least distance required for two points to be termed a neighbor): 0.046. To determine it, the distance between each data point and its nearest neighbor is calculated using the Nearest Neighbors method. Subsequently, the distances are sorted and plotted (Figure 13a). Epsilon is defined as the maximum curvature value of the resulting graph.

The clustering result for a 10% subset of Duke cases is illustrated in Figure 13b. Out of the total Duke data, the algorithm classified 510 cases, which corresponds to 6.59% of the dataset, as out-of-distribution (OOD).



(a) (b)

Figure 13: (a) Data probability curve employed to determine epsilon value. The most significant deviation is observed at approximately 0.046. (b) DBSCAN clustering technique's outcome yielded a distinction between in-distribution cases (blue) and OOD cases (red).

Figure 14 highlights several cases classified as out-of-distribution (OOD). Upon conducting a manual inspection of 50 of those cases, it was found that 38 of them exhibited clear reasons for belonging to this group.

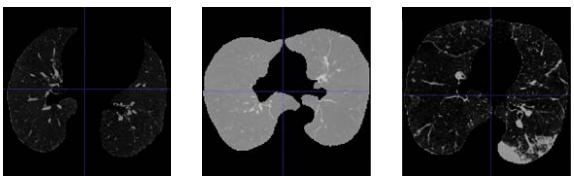
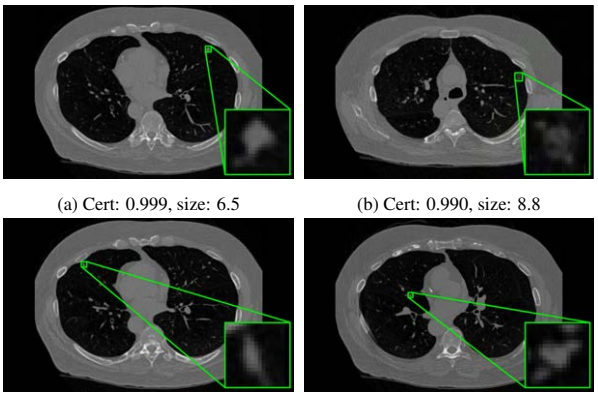


Figure 14: Left image depicts the center case of the training distribution. The middle image shows a case classified as OOD due to a difference in the reconstruction method. Finally, the right image represents an OOD case attributed to variations in patient anatomy/pathology.



(a) Cert: 0.999, size: 6.5 (b) Cert: 0.990, size: 8.8

(c) Cert: 0.748, size: 5.6 (d) Cert: 0.137, size: 6.4

Figure 15: Detected nodules in a Duke case with certainty scores and the length in mm of the largest bounding box axis. Each detection is presented in its corresponding center slice.

5.3. Transfer Learning

Using transfer learning, we incorporate Duke Data into the model trained on LIDC data. In Appendix 8, detailed information about the training and validation pipeline can be found in Figure 19. During training, the model learns to classify the anchor boxes into object or non-object and predict accurate bounding box coordinates for the objects. Once the model is trained, it can be used for inference. The model predicts the probabilities of each anchor box containing an object and refines the bounding box coordinates if an object is detected.

The model is trained and validated on the LIDC/IDLR dataset using K-fold cross-validation. The validation results demonstrate an average recall (AR) of 0.99 and an average precision (AP) of 0.858 using an Intersection over Union (IoU) threshold of 0.1. Using different IoU values (0.01, 0.1, and 0.5), we obtain a mean average recall (mAR) of 0.998 and a mean average precision (mAP) of 0.852.

Figure 15 displays several identified nodules within a single Duke case.

In order to investigate the distribution characteristics of the detected nodules, a comprehensive analysis was conducted by constructing cumulative histograms for the bounding box scores and major axis lengths. This analytical approach, depicted in Figure 16, serves as a crucial visual tool for understanding the underlying patterns and tendencies within the dataset and detection algorithm.

By examining the cumulative histograms, it is observed that approximately 75% of the lesions exhibit significant uncertainties, falling within the (0-0.2) and (0.9-1) ranges. Furthermore, approximately 50% of the detected lesions have a major axis length of less than 5 mm. Additionally, a substantial majority of 90% possess a major axis length below the 10 mm threshold.

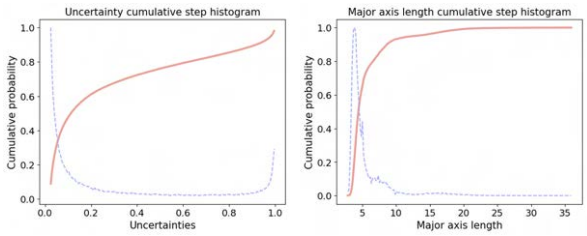


Figure 16: Cumulative Step Histograms of detected nodule probability based on uncertainty score and major axis length. The red curve represents the probability, while the blue one shows its first derivative.

5.4. Active Learning

Bellow, we analyze the results for each AL strategy.

AL Strategy 1: Uncertainty Sampling: For every epoch, we calculate the patient-level certainty for those cases. This is done by filtering and averaging the considered lesions' certainties. Among the total of 95,556 lesion candidates detected in Duke Data, 46% of them were excluded based on the low-score criteria, 17% were discarded due to their localization outside the lung region, and 1.96% were eliminated based on their size. 906 scans were pushed to the certain pool (weak labels) and 906 to the uncertain pool.

AL Strategy 2: Representativeness: From 906 uncertain scans using our previously defined representativeness (recency) criteria, we reduced them to 517 cases. Those are the ones that will finally be labeled.

AL Strategy 3: Pseudo-Labelling: Among the 906 scans within the certain pool, the average pre-filtered patient-level certainty was determined to be 0.386 ± 0.395. However, the subsequent application of more stringent filtering measures resulted in an increased mean score of 0.815 ± 0.207.

In conjunction with the LIDC cases, the acquired pseudo-labels were utilized for training the model. The results of the training and validation procedures are presented in Figure 17 and Table 3, respectively. We can observe how the use of pseudo-labeling considerably increases the model's performance. Furthermore, incorporating two weak labels per active learning epoch yields better results compared to using only one weak label.

	AP	AR	mAP	mAR
Weak labels	**0.9234**	0.9969	**0.9161**	0.9934
½ weak labels	0.8898	**0.9989**	0.8844	**0.9987**
No weak labels	0.8331	0.9836	0.8329	0.9836

Table 3: Validation metrics using the three data configurations. "AP" (Average Precision) quantifies detection model quality by calculating the area under the precision-recall curve. "AR" (Average Recall) measures the true positive rate. Both metrics are computed at IoU=0.1, determining the required overlap for valid detection. Varying IoU provides insights into localization accuracy at different levels. "mAP" and "mAR" denote mean values across 0.01, 0.1, and 0.5 IoU.

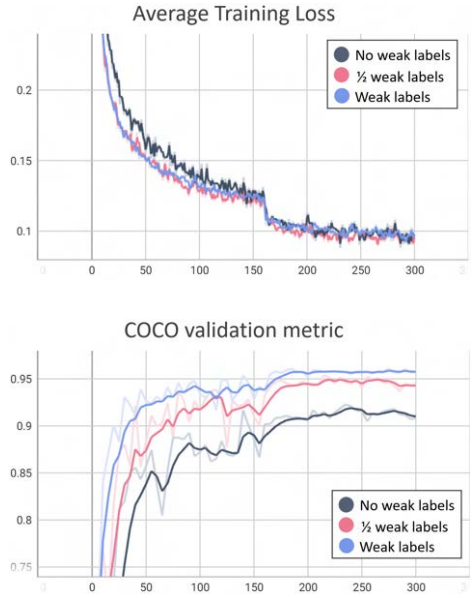


Figure 17: Training and validation curves for three distinct data configurations: (grey) "No weak labels" - LIDC cases exclusively, (pink) "Weak Labels" - LIDC cases combined with 906 Duke pseudo-labels, 2 pseudo-labels per AL epoch. (Blue) "½ Weak Labels" - LIDC cases combined with 453 Duke pseudo-labels, 1 pseudo-labels per AL epoch. The training loss depicted represents the average of cross entropy and focal loss. The validation metric displayed is COCO val.

5.5. The impact of factors on performance

Figure 18 illustrates the relationship between the analyzed factors and the study metrics. We can see that for most of the studied factors, the top and bottom groups overlap, showing that the model is robust to those factors. Nevertheless, we can find differences between groups, indicating worse model performance at the OOD organ level. Table 4 displays the mean percentage differences between the top and bottom mean for each factor. The top group for entropy, delentropy, GLCM, and noise corresponds to the 5% of the scans with the highest metric values. Conversely, for OOD, the top cases indicate that 5% of the data that is farthest from the training distribution. Regarding recency, the top cases represent the most recent instances, while the top previous cancer refers to patients that had it before.

Factor	Detection score	Included lesions
Entropy	2.04	3.88
Delentropy	2.65	0.39
GLCM	1.36	0.67
Noise	2.59	2.27
OOD Organ	**5.68**	4.16
OOD Patient	4.70	3.14
Prosthesis	0.46	1.92
Recency	1.39	0.85
Previous cancer	3.49	**4.61**

Table 4: Significant variations across factors observed between the highest and lowest groups.

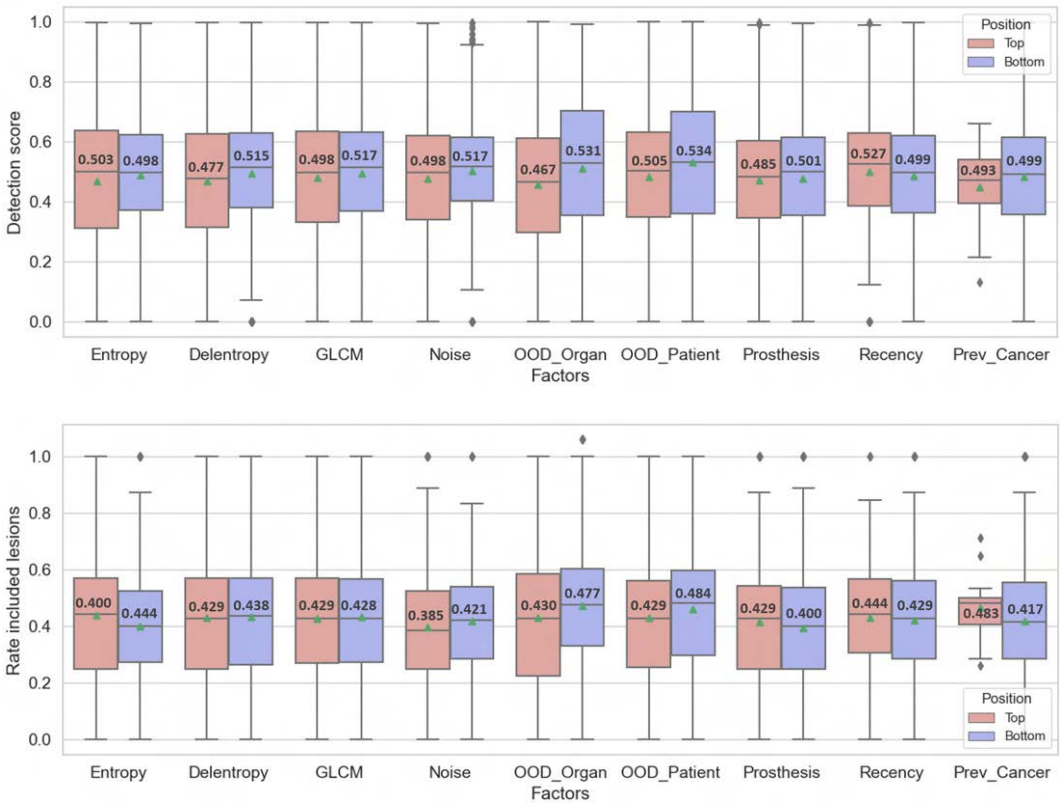


Figure 18: Box-and-whisker plots illustrating the relationship between uncertainty and the rate of included lesions across different factors. The median value is displayed in each plot, and the mean is represented with a green triangle. Sub-figure (a) depicts the uncertainty for the top and bottom cases per analyzed factor. The case-level uncertainty is derived by averaging the filtered lesion-level uncertainties, utilizing the criteria elucidated in Section 4.2. While, (b) illustrates the rate of included lesions for the top and bottom cases per analyzed factor. The filtering process excludes detections that are unlikely to be nodules. To calculate this rate, we divide the number of lesions considered after filtering by the total number of lesions detected by the algorithm in each scan.

6. Discussion and Interpretation

6.1. Dataset characterization

Noise levels in both datasets, Duke Dataset and LIDC, are quite similar. Several factors, including acquisition/reconstruction parameters, total attenuation, and tissue density, may contribute to image noise. Duke Dataset has higher noise. That could be because all its cases are low-dose CT, whereas LIDC encompasses a mixture of standard and low-dose cases. On the other hand, Duke cases are newer and are mostly iterative reconstructed that would apply nonlinear noise reduction. This may be why the 2 datasets seem similar. Furthermore, the increased standard deviation observed in the Duke data likely stems from its broader range of images collected over an extended timeframe.

In terms of texture complexity, LUNA16 is considered the most challenging dataset for classification algorithms. It has the highest values for Shannon Entropy and Delentropy. Shannon Entropy focuses on individual pixel values, while Delentropy captures non-local information and represents higher-order image structures. However, the most significant difference between the datasets is observed at the global level. Duke Dataset has higher GLCM-based entropy, indicating that its volumes contain more diverse texture patterns due to the use of various detectors and protocols. This results in a greater disparity in entropy values for global measures.

For a deep learning network, higher pixel entropy suggests that the initial layers, which interpret pixel distributions, will struggle to learn. This can lead to issues such as biased weights and longer convergence times. On the other hand, higher delentropy and GLCM values imply that the network will face challenges in learning higher-order features in the middle or later layers.

Our analysis reveals minimal disparities in noise and textures across the datasets, thus substantiating the homogeneity of semantic data.

6.2. Out-of-distribution analysis

Focusing on the patient level, for our configuration, the proposed distance-based method (histogram features) works better than the reconstructed one (VAE). However, this first approach is dependent on the assumption of semantic homogeneity within the data, which cannot be assumed across all scenarios. (Frolova et al., 2022) also reports better results using histogram features than other deep learning approaches.

On the other hand, VAE can handle nonlinearities in data and doesn't require semantic homogeneities. Nevertheless, an evident limitation lies in their inherent dependence on the expressive capacity, i.e., the size and configuration of the latent space for effectively reconstructing anomalies. Consequently, reconstruction-based techniques continue to yield notable performance scores in unsupervised tasks, primarily due to their ability to compensate for deficiencies to some extent by fine-tuning the model architecture specifically tailored to the given task. However, it is important to note that task-specific hyperparameter optimization deviates from the principle of assumption-free anomaly detection, which prioritizes a more agnostic approach.

Focusing on the organ's distribution, we can find anatomy shifts and also different image reconstruction protocols. Although the accuracy at this level is lower, the classification becomes more challenging due to data shifts limited to within the organ.

The results prove the importance of making the OOD analysis into different levels since anatomical and pathological shifts are not detected at the patient level.

6.3. Active Learning

Going from the lesion-certainties to the patient-certainties is dependent on the filtering criteria chosen. 17% of the detected lesions are outside the dilated lung mask. To avoid that and consequently decrease the false positive rate (FPR), the input of the network could be the masked volumes. Future work could include this experiment to ensure there are no contextual misunderstandings, broken spatial relationships, or incomplete information that worsen the model training. The higher rate of discarded lesions is due to the detection scores.

We cannot evaluate the proposed AL method completely until the data is manually labeled. But regarding the pseudo-labeling, we can conclude that is a promising strategy in the realm of active learning, although its added benefits compared to uncertainty sampling should be lower because the model already knows how to detect those cases. Adding more data from another dataset will increase the model's generalization capability.

As inspired by (Gorriz et al., 2017), the approach presented here entails a straightforward strategy that could be further refined by updating pseudo-labels as the model improves during training or returning samples to the unlabelled pool when they become uncertain. Recent advancements in this field have introduced intriguing alternative approaches, such as the reinforcement learning method proposed by (Park et al., 2018). This particular approach claims significant reductions in annotation efforts for the challenging task of lung nodule detection in chest X-rays. These novel techniques hold great potential for enhancing the efficiency and effectiveness of annotation processes in various domains.

6.4. How does my model see my data?

Beginning with the examination of textures, there are no significant differences observed between the groups. When we shift our focus to the detection scores, we find that higher levels of detail (indicated by larger entropy values) result in lower scores across all three metrics. As mentioned earlier, these high-detail cases are more challenging for the model to learn from since it has primarily been trained on cases with lower levels of detail (LIDC cases). In terms of the rate of included lesions, the Shannon Entropy metric shows the most notable difference. We notice that as the SE value increases, the rate of included lesions also increases. This can be attributed to the fact that SE relies on individual pixels, and when the data contains more localized information, the model tends to detect fewer false positives.

Studying the noise, we consider as top group the cases with higher noise levels. These cases lead to decreased model confidence and an increase in the false positive rate (FPR). Furthermore, the detection scores exhibit a greater dispersion within this group. Noise negatively impacts our specific configuration.

Moving on to OOD organ detection cases, we define the top cases as those that deviate the most from the training distribution based solely on lung volumes. These cases display the lowest detection scores, meaning that those cases confuse our model the most. Additionally, the rate of included lesions shows the most significant difference between top-bottom groups. Some cases in this group exhibit additional pathologies, such as pneumonia, which further complicates accurate prediction and contributes to an increased FPR.

When considering patient-level analysis, we arrive at similar conclusions but with smaller differences between groups. This indicates that the deviation in data within the lungs has a more pronounced impact on our model compared to deviations outside of the lungs. As previously mentioned, the accuracy of OOD organ analysis is lower than that of patient-level analysis. However, the ones spotted at the organ level present a higher impact. That means that even if there is no prior reason for considering a case out-of-distribution, our method is finding data shifts that deteriorate the predictions.

Patients with prostheses introduce some level of uncertainty, although not as much as observed in any OOD analysis. This suggests that prostheses do not significantly affect our model.

Regarding the recency factor, we consider the top group the most recent scans per patient. We find that the model is more confident with the last scans. Risk patients are usually followed-up, where the lesions may increase their size. As it varies the model performance, it supposes a good criterion for representativeness filtering in the AL approach.

Previously cancer-treated patients confuse the model. Nevertheless, as the group of patients with previous

surgeries is less populated, these conclusions are not representative. A bigger cohort should be studied.

Despite the model's susceptibility to the OOD factors, the differences between groups are not significantly pronounced for the majority of them. This observation demonstrates the considerable robustness of the studied model. However, it is crucial to acknowledge that alternative models or datasets could display less overlap between groups, resulting in more discernible performance differences.

The robustness of our model can be attributed to several factors. Firstly, RetinaNet adopts a two-stage architecture consisting of a Region Proposal Network (RPN) and a classification subnet. This design enhances robustness by improving localization accuracy. Additionally, by integrating features extracted at different levels of the multi-scale feature pyramid, RetinaNet effectively captures both fine-grained details and high-level semantic information, thereby handling variations more effectively and increasing the model's resistance to noise and texture shifts. Another key aspect contributing to the model's robustness is the consideration of multiple anchor boxes for each object instance. Even if noise affects some of the anchor boxes, the model can still rely on others to accurately detect the object. Furthermore, the incorporation of trained techniques such as dropout, weight decay, and batch normalization aids in reducing the model's dependence on specific training samples and enhances its resilience to out-of-distribution cases, promoting generalization.

Future work could analyze which labeling subset improves more the model performance, whether the one proposed by an Active Learning algorithm or the one identified through OOD organ-level analysis. Furthermore, considering that organ-level OOD cases can cause confusion for our model, we have two options to ensure reliable outcomes: either manually label these cases to train the model on accurate detection or exclude them from the study altogether.

7. Conclusions

In this study, we have successfully demonstrated the efficacy of OOD analysis in providing valuable insights into the underlying mechanisms by which a model perceives and interprets data. Furthermore, we have introduced a novel approach wherein OOD factors are employed as a means of explainability, a concept that has not been previously proposed in the existing literature. Additionally, we have devised a comprehensive pipeline highly recommended to follow when incorporating new unlabelled datasets into a study, as doing so ensures the model's ability to make robust and trustworthy predictions. Due to the absence of ground truth in the data, which hinders the evaluation of the model's data affection, we propose an evaluation methodology grounded in filtering techniques and detection scores. Furthermore, our pipeline incorporates an Active Learning (AL) approach that employs a dual-criteria framework and integrates weak labels, presenting novel strategies for detection to minimize the amount of labeled data needed. The utilization of this pipeline and the integration of OOD factors as an explainability approach hold significant implications for advancing the field of machine learning interpretability. By providing a comprehensive framework for identifying and understanding data abnormalities, enhancing model transparency and robustness.

Acknowledgments

I would like to express my deepest appreciation to Dr. Joseph Y. Lo for his exceptional supervision and support throughout this research project. His expertise and guidance have been invaluable in shaping the study and pushing it toward excellence. I am also grateful to the Duke CVIT lab for hosting my project and providing the necessary resources. Additionally, I extend my heartfelt thanks to my family for their unwavering support and to MaIA Master program for equipping me with the tools necessary to excel in this endeavor.

References

Alain, G., Bengio, Y., 2014. What regularized auto-encoders learn from the data generating distribution. arXiv:1211.4246.

Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., et al., 2021. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiology: Artificial Intelligence 3, e200267.

Bank, D., Greenfeld, D., Hyams, G., 2018. Improved training for self training by confidence assessments, in: Intelligent Computing: Proceedings of the 2018 Computing Conference, Volume 1, Springer. pp. 163–173.

Beluch, W.H., Genewein, T., Nürnberger, A., Köhler, J.M., 2018. The power of ensembles for active learning in image classification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 9368–9377.

Bengar, J.Z., Gonzalez-Garcia, A., Villalonga, G., Raducanu, B., Aghdam, H.H., Mozerov, M., Lopez, A.M., Van de Weijer, J., 2019. Temporal coherence for active learning in videos, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), IEEE. pp. 914–923.

Bengar, J.Z., Raducanu, B., van de Weijer, J., 2021. When deep learners change their mind: Learning dynamics for active learning, in: Computer Analysis of Images and Patterns: 19th International Conference, CAIP 2021, Virtual Event, September 28–30, 2021, Proceedings, Part I, Springer. pp. 403–413.

Berger, C., Paschali, M., Glocker, B., Kamnitsas, K., 2021. Confidence-based out-of-distribution detection: a comparative study and analysis, in: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3, Springer. pp. 122–132.

Bongartz, G., 1999. European guidelines on quality criteria for computed tomography. http://www. drs. dk/guidelines/ct/quality/htmlindex. htm .

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A., 2018. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians 68, 394–424.

Brust, C.A., Käding, C., Denzler, J., 2018. Active learning for deep object detection. arXiv:1809.09875.

Cai, W., Zhang, Y., Zhou, S., Wang, W., Ding, C., Gu, X., 2014. Active learning for support vector machines with maximum model change, in: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I 14, Springer. pp. 211–226.

Chen, X., Konukoglu, E., 2018. Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. arXiv preprint arXiv:1806.04972 .

von Eschenbach, W.J., 2021. Transparency and the black box problem: Why we do not trust ai. Philosophy & Technology 34, 1607–1622.

Fong, R.C., Vedaldi, A., 2017. Interpretable explanations of black boxes by meaningful perturbation, in: Proceedings of the IEEE international conference on computer vision, pp. 3429–3437.

Frolova, D., Vasiliuk, A., Belyaev, M., Shirokikh, B., 2022. Solving sample-level out-of-distribution detection on 3d medical images. arXiv preprint arXiv:2212.06506 .

Gal, Y., Islam, R., Ghahramani, Z., 2017. Deep bayesian active learning with image data, in: International conference on machine learning, PMLR. pp. 1183–1192.

Golestaneh, S.A., Kitani, K.M., 2020. Importance of self-consistency in active learning for semantic segmentation. arXiv preprint arXiv:2008.01860 .

Gorriz, M., Carlier, A., Faure, E., i Nieto, X.G., 2017. Cost-effective active learning for melanoma segmentation. arXiv:1711.09168.

Gu, Y., Chi, J., Liu, J., Yang, L., Zhang, B., Yu, D., Zhao, Y., Lu, X., 2021. A survey of computer-aided diagnosis of lung nodules from ct scans using deep learning. Computers in biology and medicine 137, 104806.

Guo, Y., 2010. Active instance sampling via matrix partition. Advances in Neural Information Processing Systems 23.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Hendrycks, D., Gimpel, K., 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136 .

Hsu, Y.C., Shen, Y., Jin, H., Kira, Z., 2020. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10951–10960.

Huang, S.J., Jin, R., Zhou, Z.H., 2010. Active learning by querying informative and representative examples. Advances in neural information processing systems 23.

Käding, C., Rodner, E., Freytag, A., Mothes, O., Barz, B., Denzler, J., AG, C.Z., 2018. Active learning for regression tasks with expected model output changes., in: BMVC, p. 103.

Kao, C.C., Lee, T.Y., Sen, P., Liu, M.Y., 2019. Localization-aware active learning for object detection, in: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part VI 14, Springer. pp. 506–522.

Karimi, D., Gholipour, A., 2022. Improving calibration and out-of-distribution detection in deep models for medical image segmentation. IEEE Transactions on Artificial Intelligence .

Larkin, K.G., 2016. Reflections on shannon information: In search of a natural information-entropy for images. arXiv preprint arXiv:1609.01117 .

Lee, D.H., et al., 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks, in: Workshop on challenges in representation learning, ICML, p. 896.

Liang, S., Li, Y., Srikant, R., 2017. Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv preprint arXiv:1706.02690 .

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. arXiv:1612.03144.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2018. Focal loss for dense object detection. arXiv:1708.02002.

Makaju, S., Prasad, P., Alsadoon, A., Singh, A., Elchouemi, A., 2018. Lung cancer detection using ct scan images. Procedia Computer Science 125, 107–114.

Pacheco, A.G., Sastry, C.S., Trappenberg, T., Oore, S., Krohling, R.A., 2020. On out-of-distribution detection algorithms with deep neural skin cancer classifiers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 732–733.

Park, S., Hwang, W., Jung, K.H., 2018. Integrating reinforcement learning to self training for pulmonary nodule segmentation in chest x-rays. arXiv preprint arXiv:1811.08840 .

Rahane, A.A., Subramanian, A., 2020. Measures of complexity for large scale image datasets, in: 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), IEEE. pp. 282–287.

Ronald J. Brachman, P.S., . Synthesis lectures on artificial intelligence and machine learning .

Saito, P.T., Suzuki, C.T., Gomes, J.F., de Rezende, P.J., Falcao, A.X., 2015. Robust active learning for the diagnosis of parasites. Pattern Recognition 48, 3572–3583.

Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G., 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery, in: Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings, Springer. pp. 146–157.

Sener, O., Savarese, S., 2017. Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489 .

Setio, A.A.A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., Van Riel, S.J., Wille, M.M.W., Naqibullah, M., Sánchez, C.I., Van Ginneken, B., 2016. Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. IEEE transactions on medical imaging 35, 1160–1169.

Shyu, M.L., Chen, S.C., Sarinnapakorn, K., Chang, L., 2003. A novel anomaly detection scheme based on principal component classifier. Technical Report. Miami Univ Coral Gables Fl Dept of Electrical and Computer Engineering.

Team, N.L.S.T.R., 2011. The national lung screening trial: overview and study design. Radiology 258, 243–253.

Wang, J., Wen, S., Chen, K., Yu, J., Zhou, X., Gao, P., Li, C., Xie, G., 2020. Semi-supervised active learning for instance segmentation via scoring predictions. arXiv preprint arXiv:2012.04829 .

Wasserthal, J., Meyer, M., Breit, H.C., Cyriac, J., Yang, S., Segeroth, M., 2022. Totalsegmentator: robust segmentation of 104 anatomical structures in ct images. arXiv preprint arXiv:2208.05868 .

Wisselink, H.J., Pelgrim, G.J., Rook, M., Dudurych, I., van den Berge, M., de Bock, G.H., Vliegenthart, R., 2021. Improved precision of noise estimation in ct with a volume-based approach. European Radiology Experimental 5, 1–7.

Xu-Darme, R., Girard-Satabin, J., Hond, D., Incorvaia, G., Chihani, Z., 2023. Interpretable out-of-distribution detection using pattern identification. arXiv preprint arXiv:2302.10303 .

Yang, Y., Loog, M., 2018. A variance maximization criterion for active learning. Pattern Recognition 78, 358–370.

Yang, Y., Ma, Z., Nie, F., Chang, X., Hauptmann, A.G., 2015. Multi-class active learning by uncertainty sampling with diversity maximization. International Journal of Computer Vision 113, 113–127.

Zimmerer, D., Kohl, S.A.A., Petersen, J., Isensee, F., Maier-Hein, K.H., 2018. Context-encoding variational autoencoder for unsupervised anomaly detection. arXiv:1812.05941.

Zimmerer, D., Petersen, J., Kohl, S.A.A., Maier-Hein, K.H., 2019. A case for the score: Identifying image anomalies using variational autoencoder gradients. arXiv:1912.00003.

Zotova, D., Lisowska, A., Anderson, O., Dilys, V., O'Neil, A., 2019. Comparison of active learning strategies applied to lung nodule segmentation in ct scans, Springer. pp. 3–12.

8. Appendix

Figure 19 shows the training and inference of the model. The anchor generator is responsible for generating a set of anchor boxes at different scales and aspect ratios across the image. These anchor boxes act as reference bounding boxes that the model will use to make predictions. The anchor generator typically generates anchor boxes at multiple spatial locations across different feature levels in the FPN.

The vanilla/ATSS matcher is used during training to assign positive and negative samples to each anchor box. It helps to determine which anchor boxes are considered positives (containing objects of interest) and which are considered negatives (background or non-object regions). The ATSS matcher adapts the assignment strategy based on the distribution of object sizes within the dataset, which helps handle imbalanced data.

RetinaNet's training involves the following steps:

- Anchor Generation: The anchor generator generates a set of anchor boxes across the feature pyramid. Each anchor box is associated with a specific spatial location and aspect ratio.
- Matching Anchors: The vanilla ATSS matcher assigns positive and negative labels to anchor boxes by evaluating their overlap with ground truth bounding boxes. The ATSS matcher adapts the matching threshold dynamically, taking into consideration the distribution of object sizes in the dataset.
- Loss Calculation: The focal loss down-weights the loss contribution from easy negative samples, which helps in handling the imbalance between background and foreground regions. Loss is calculated for both the classification (object/non-object) and bounding box regression tasks.
- Training: The total focal loss of an image is calculated by summing the focal loss over all approximately 100,000 anchors. This sum is then normalized by the number of anchors that have been assigned to a ground-truth box. Network initialization is very important. A prior probability of 0.01 is assumed for all anchor boxes and assigned that bias to the last Conv. Classification subnet layers.
- Inference: The network limits the decoding of box predictions to a maximum of 1,000 top-scoring predictions per Feature Pyramid Network (FPN) level. This is done after setting a threshold of 0.05 for the detector confidence. Afterward, the top predictions from all levels are combined, and a technique called non-maximum suppression (NMS) is applied with a threshold of 0.5. This process results in the final detections.

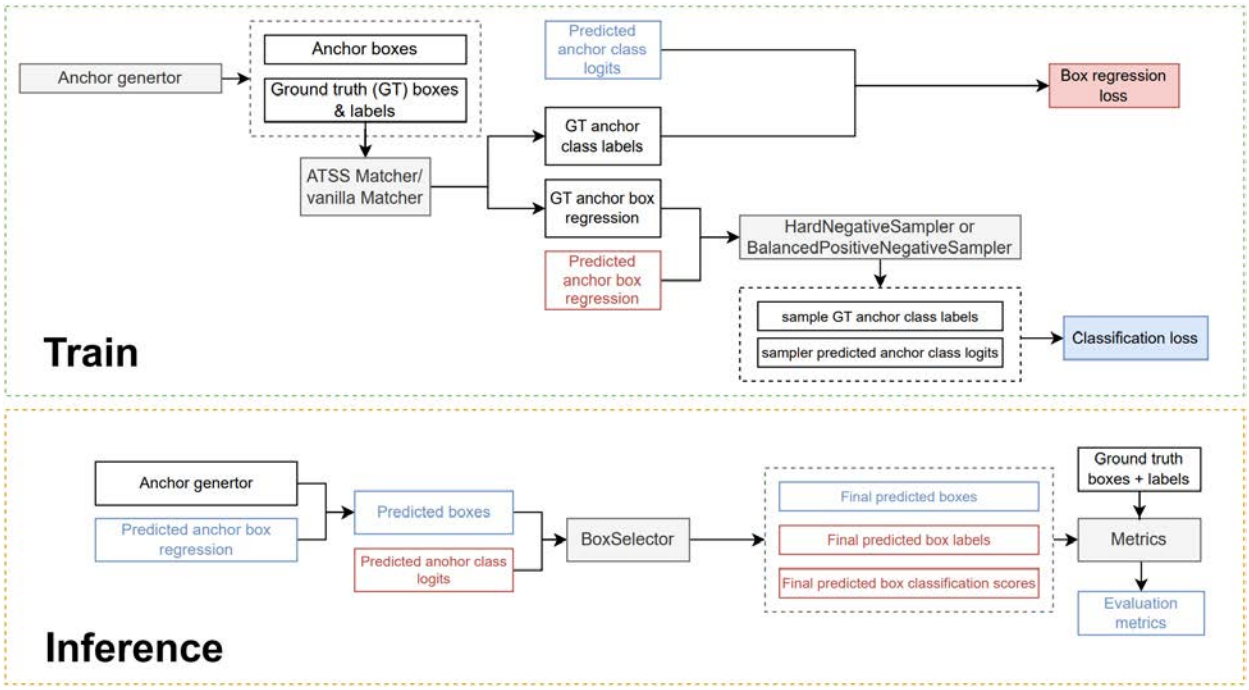


Figure 19: Pipeline for training and inference using MONAI's model.

Medical Imaging and Applications

Master Thesis, June 2023

Inferring morphological patterns of EGFR gene mutation from lung cancer tissues using large scale architectures

Nisma Amjad[1], Natalia Valderrama[1], Nathan Vincon[1]

[a] *Ummon HealthTech, Dijon, France*

Abstract

Deep learning methods combined with digital pathology have the potential to generate new biomarkers for diagnostic, prognostic, and theranostic purposes. However, their lack of transparency hinders their acceptance in clinical practice. This study aims to establish and compare different methods from the literature for extracting interpretable features from a black-box deep learning model. The research leverages Whole Slide Images (WSI) and machine learning techniques to enhance the understanding of Epidermal Growth Factor Receptor (EGFR) mutations in lung adenocarcinoma (LUAD). Using AI explainability, a machine learning model is trained on the TCGA LUAD dataset to gain insights into the underlying factors and patterns contributing to EGFR mutation detection. A deep learning (DL) model is developed to predict EGFR mutation status in LUAD samples. The DL model's prediction scores are used as labels for the machine learning models. Regression and classification models are applied to improve interpretability, utilizing selected features from pixel intensity statistics, color features, texture features, and shape features. The performance of different machine learning models is evaluated using 5-fold cross-validation and an 80:20 split. The linear Support Vector Regression (SVR) model performs better in cross-validation, while the Random Forest (RF) model outperforms others in the 80:20 split. The RF model achieves higher R-squared scores for both 144 and 54 selected features. Additionally, ROC curve analysis shows that the RF model exhibits better discriminative ability, with a maximum AUC of 0.667. These findings contribute to the development of interpretable deep learning models and improve the accuracy of EGFR mutation analysis in LUAD.

Keywords: Epidermal Growth Factor Receptor (EGFR) mutations, Lung adenocarcinoma (LUAD), Machine learning, Deep learning, Convolutional neural network (CNN), Support Vector Regression (SVR), Random Forest (RF), Feature selection, Cross-validation, Interpretability, Performance evaluation

1. Introduction

According to recent data from the World Health Organization (WHO) [1], the global burden of cancer continues to escalate significantly. As of 2020, an estimated 20 million new cancer cases were reported worldwide. This places cancer as the second leading cause of death globally, following cardiovascular disease, according to the Global Burden of Disease Cancer Collaboration. It is projected that one in five men and one in six women will develop cancer during their lifetime. These statistics highlight the alarming prevalence of this disease across genders. Furthermore, the WHO predicts a staggering 60 percent increase in cancer cases by the year 2040, with an estimated 9.6 million individuals succumbing to the increasing effects of cancer. Estimated number of incident cases and deaths World due to cancer is shown in figure 1.

Among the various types of cancer, lung cancer stands out as particularly prevalent and fatal. Annually, over a million individuals from all corners of the world lose their lives due to the impact of lung cancer (Ferlay et al., 2015).

Lung cancer is customarily separated into two principal categories based on cellular composition: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) (Travis et al., 2015). Non-small cell lung can-

[1] `https://www.iarc.who.int`

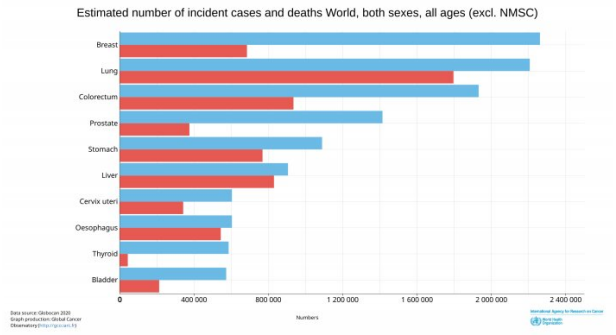


Figure 1: Estimated number of incident cases and deaths World due to cancer till year 2020, both sexes, all ages (excl. NMSC). Here red represents mortality rate and blue represents incidence rate(WHO)

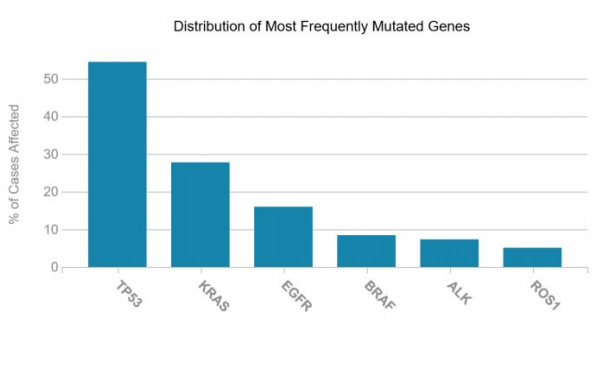


Figure 2: Distribution of most frequently mutated genes incl. TP53, KRAS, EGFR, BRAF, ALK and ROS1 (2023.)

cer (NSCLC), constituting approximately 85% of all lung cancer cases, encompasses a heterogeneous cluster of subtypes characterized by distinct cellular characteristics (Couraud et al., 2012; Howlander et al., 2019). Among these subtypes are adenocarcinoma (LUAD), squamous cell carcinoma (LUSC), and large-cell carcinoma.

Accurate identification and classification of each NSCLC subtype, as well as the determination of the disease stage, are crucial for optimal management and personalized treatment decisions. Diagnosis typically involves the collection of lung tissue samples through biopsy procedures targeting the affected regions. Subsequently, targeted treatment approaches are tailored based on factors such as subtype classification, staging status, and the presence of specific genetic mutations, aiming to effectively combat the disease within the context of each individual case (Chan and Hughes, 2015).

Among the genetic alterations observed in NSCLC, several key mutations have been identified, including Epidermal Growth Factor Receptor (EGFR) gene alterations, Anaplastic Lymphoma Kinase (ALK) and ROS1 gene rearrangements, BRAF gene modifications, and NTRK gene fusions. (Ettinger et al., 2018). Mutations in the BRAF gene, found in 2-4% of NSCLCs, can be targeted with BRAF inhibitors, improving treatment outcomes for affected patients and advancing personalized medicine in lung cancer (Baik et al., 2017). The study by Shaw et al. (2013) says that nearly 5% of NSCLCs have ALK gene rearrangements, most commonly leading to fusion with EML4. About 20% of LUAD patients have EGFR mutations, while KRAS and TP53 mutations are very common about 25% and 50% respectively. These mutations have proved particularly challenging to target however by identifying these mutations healthcare professionals can select treatment options that specifically target the molecular characteristics of the tumor, thereby maximizing the effectiveness of the treatment and improving patient outcomes(Chan and Hughes, 2015) (Terra et al., 2016).

Differentiating between mutations can be challenging, particularly in poorly differentiated tumors where additional studies are necessary to ensure accurate categorization. Recently, lung cancer whole-slide image analysis, utilizing deep convolutional neural networks (DCNNs), has proven valuable in survival prognosis and classification. However, a limitation is that these models lack interpretability. Machine learning (ML) models can leverage histological features to identify patterns and biomarkers associated with disease progression and patient outcomes (Levitsky et al., 2019) (Lai et al., 2020). This automated scrutiny of slide images contributes to the precision and reliability of information for improved decision-making in the management of NSCLC. (Luo et al., 2017)(Yu et al., 2016).

This study presents a novel approach to infer Epidermal Growth Factor Receptor (EGFR) mutations through AI explainability using a machine learning (ML) model applied to the The Cancer Genome Atlas (TCGA) LUAD dataset. The study focuses on utilizing whole slide images (WSI) from the TCGA dataset to detect EGFR mutations by leveraging a convolutional neural network (CNN) based architecture. In this study, machine learning plays a crucial role in providing explainability and interpretability for the predictions made by the deep neural network. The primary objective is to understand the factors and patterns contributing to the detection of Epidermal Growth Factor Receptor (EGFR) mutations in lung adenocarcinoma (LUAD) samples.

By analyzing the machine learning models through features, we aim to gain insights into the underlying factors that drive the EGFR mutation detection task. These features capture relevant information from the input data, such as visual characteristics of tumor tissues in whole slide images (WSI) from the The Cancer Genome Atlas (TCGA) LUAD dataset. By examining these features, the machine learning models can identify and learn patterns associated with EGFR mutations. By gaining insights into the deep learning model's decision-making process and identifying the patterns it learns, this approach enhances our understanding of

how the model detects EGFR mutations. This knowledge can have significant implications for improving diagnostic and treatment strategies for patients with LUAD. It allows for the development of more targeted and effective approaches by leveraging the identified patterns and factors associated with EGFR mutations.

Overall, the role of machine learning in this study is to provide explainability to the predictions of the deep neural network, enabling a deeper understanding of EGFR mutation detection in LUAD. The analysis of features and patterns contributes to a comprehensive interpretation of the model's performance, facilitating advancements in the field of lung cancer diagnosis and treatment.

2. State of the art

Lung cancer denotes a predominant origin of cancerous fatalities around the globe. It constitutes a varied affliction, categorized into manifold subsets founded on microscopic traits and molecular deviations. Ascertainment and comprehension of these subsets are imperative for precise diagnosis, prognosis, and the evolution of tailored therapies. The most frequently detected lung cancer around the world is the primary cause of cancer mortality for both genders – approximated at 1.6 million expirations in 2018 (Bray et al., 2018).

The first classification of non-small cell lung cancer (NSCLC) are lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) (Molina et al., 2008). Correct. NSCLC (non-small cell lung cancer) predictions often take into account various factors, including tobacco smoking history, family history, and occupational exposures. These factors play a significant role in assessing the risk and prognosis of NSCLC (Cruz and Wishart, 2006)..

In addition to these factors, biomarkers such as TP53 mutations, KRAS, ALK, and EGFR have been identified as important in the context of smoking-related lung cancer. These biomarkers can influence the survival rates, treatment response, and overall prognosis of individuals with NSCLC. They provide valuable insights into the underlying molecular characteristics of the cancer and help guide personalized treatment approaches. Furthermore, these biomarkers should fit together to produce an accurate and robust prediction of a patient's prognosis (Camidge et al., 2018). The first molecular changes to confer lung cancer sensitivity to targeted therapies are EGFR mutations (Shi et al., 2014).

Tissue biopsy or surgery remains the ultimate method for diagnosing tumors (Schabath and Cote, 2019). However, these procedures come with significant risks, including surgery-related complications and time costs. To address the need for a non-invasive alternative, the integration of genomic, proteomic, and imaging technologies has emerged as powerful diagnostic tools, providing more specific molecular information about tumors and patients (Vargas and Harris, 2016). Recent research has focused on correlating radiomic biomarkers, obtained by analyzing quantification features in radiological images, with lung cancer prognosis and mutation status (Arimura et al., 2019) (Thawani et al., 2018). To leverage these advancements, deep convolutional neural networks (DCNNs) have been applied to whole slide images (WSI) using large pre-trained networks as feature extractors (Källén et al., 2016). The use of DCNNs was first proposed for histology image analysis in Cireşan et al. (2013), where the authors were able to train a model for mitosis detection in H&E biopsy images (Whole Slide Images - WSI). A similar technique was applied for invasive ductal carcinoma detection in Cruz-Roa et al. (2014).

Notably, deep learning techniques have shown promising results in classifying and predicting mutations based on histopathology images (Fu et al., 2020) (Kather et al., 2020). As H&E (Hematoxylin and Eosin) Whole Slide Images (WSI) images provide valuable information about tissue morphology, cell architecture, and various pathological features. They are instrumental in the identification and characterization of cancerous cells, tumor grading, and understanding the tumor microenvironment. Thus, these models have the advantage of automatically extracting morphological characteristics associated with the problem to be solved without the need for human intervention to identify them. They also make it possible to identify new features unknown to human expertise that can better predict the type of tumor. Automatic feature extraction has allowed the development of new fields of research. Indeed, it has made it possible to explore new questions, such as the identification of new characteristics correlated with the chances of survival of patients, to identify cells with molecular atypia traditionally discovered using costly molecular biology methods in Sirinukunwattana et al. (2021), or to identify subgroups of patients more sensitive to certain chemotherapy molecules (Yamashita et al., 2021). As well as classification (Aresta et al., 2019) (Bandi et al., 2018), segmentation (Campello et al., 2021) (Yang et al., 2018)(Zhuang et al., 2019) and other challenges, deep learning has demonstrated state-of-the-art performance.

However, the learning time of such DCNN architectures is extremely large therefore a pretrained network is used (Song et al., 2017).

In this study, we employed machine learning techniques to capture and model the distinct morphological and molecular variations in immune patterns observed in digitized hematoxylin and eosin (H&E) images from The Cancer Genome Atlas (TCGA) Lung Adenocarcinoma (LUAD) dataset. Our approach utilized a convolutional neural network (CNN) based architecture, where EfficientNet was employed for feature extraction, followed by a multilayer perceptron for further classification.

EfficientNet, chosen as the backbone architecture for feature extraction, offers several advantages. Notably, it has demonstrated remarkable performance on the ImageNet dataset, showcasing its ability to achieve optimal results in image classification tasks. One key advantage of EfficientNet is its optimization for memory usage, making it well-suited for handling large neural networks without causing memory saturation. This feature is particularly beneficial when processing data at multiple resolutions, a characteristic that is essential when working with H&E whole-slide images (WSI). By efficiently processing data at different scales, EfficientNet allows for comprehensive analysis and characterization of immune patterns present in the digitized HE images from the TCGA LUAD dataset. (Tan and Le, 2019)

Despite the benefits and promising performance of DL models, the adoption of DL networks in healthcare has been relatively slow. One main contributing factor is the limited understanding of how AI tools function before their implementation in clinical settings (Lipton, 2017). Deep learning models are often characterized by inherent opacity, making them challenging to interpret and understand due to their complex and intricate nature. Unlike traditional object-oriented code or decision trees, comprehending the inner workings of deep learning models is not as straightforward. The lack of interpretability hinders their widespread acceptance and integration into healthcare systems (Durán and Jongsma, 2021). Furthermore, General Data Protection Regulation (GDPR) law requires a transparent description of the algorithm's decision-making process before its use for patient care (Temme, 2017).

By integrating various elements into a comprehensive model, AI tools in medical imaging can somehow assist clinicians in making informed decisions. However, the limited explanatory capabilities of these models restrict their overall usefulness. Despite providing actionable outputs, the absence of explanations hampers their potential impact in clinical settings. Therefore, efforts are needed to enhance interpretability and bridge the gap between AI outputs and clinician understanding. The interpretability of DL systems not only reveals any inherent errors within the algorithms, but also enables discovery of other key points in the imaging data. Legal and ethical requirements aside, untangling the black box nature of DL systems is vital for fostering clinical trust and troubleshooting systems. Interpretability methods can also reveal new imaging biomarkers to better understand DL models (Arrieta et al., 2020). An explainable artificial intelligence (XAI) solution provides details about its functioning to end users so theyare better understood (Arrieta et al., 2020) (Lipton, 2018) (Rudin, 2019).

Our research focuses on harnessing the power of machine learning techniques to improve the interpretability of deep learning models. By utilizing machine learning, we aim to provide explanations and insights into the decision-making process of deep learning algorithms. Machine learning techniques have demonstrated their effectiveness in predicting cancer outcomes by uncovering patterns and relationships in complex datasets. The field of cancer prognosis and prediction has witnessed wide application, with studies such as Levitsky et al. (2019) and Lai et al. (2020) exploring its potential.

In the context of histology images, (Wang et al., 2017) calculated 283 morphological features from lung adenocarcinoma (LUAD) patients' histology images from the TCGA project. They identified strong correlations between certain morphological features and patient outcomes. These features encompassed geometry, pixel intensity, and texture characteristics. Geometry features focused on cell shape, including measurements like area, perimeter, and solidity calculated from segmentation masks. Pixel intensity statistics captured cell color through metrics such as mean, standard deviation, and entropy. Texture features analyzed patterns related to protein expressions, utilizing methods like co-occurrence matrices and local binary patterns to reveal local texture patterns.

Other studies, such as Rossi et al. (2021), explored the analysis of shape and textural features in CT scans for diagnostic purposes. Their machine learning model, which considered both radiomic and clinical features, achieved a diagnostic accuracy of 88.1% in a dataset from the Cancer Imaging Archive (TCIA). Moreover, investigations into EGFR mutation status correlation with CT scan imaging phenotypes revealed significant associations. For example, Digumarthy et al. (2019) achieved an AUC of 0.79 by combining radiomic features with clinical data. Mei et al. (2018) obtained AUCs of 0.58 and 0.66 using only radiomic features and combined radiomic and clinical features, respectively. (Liu et al., 2016) showed improved predictive power by 0.67 to 0.71 using radiomic features, while Liu et al. (2016) demonstrated that combining clinical variables and CT features resulted in higher AUC values (0.78) compared to clinical variables alone. Furthermore, Yu et al. (2016) employed a combination of conventional image processing techniques and machine learning methods, such as random forest classifiers, SVM, or Naïve Bayes classifiers, to analyze lung cancer whole-slide images. Their approach achieved high accuracy achieving an Area Under the Curve (AUC) of 0.85 in distinguishing normal from tumor slides, and 0.75 in distinguishing LUAD from LUSC slides.

Overall, machine learning methods have the potential to substantially enhance the accuracy of predicting various aspects of cancer, such as susceptibility, recurrence, and mortality, as reported in the literature which are further implemented by us.

3. Material and Methods

In our approach, the whole slide images were initially divided into patches, which were then fed into a deep neural network for EGFR mutation detection. The deep neural network provided prediction scores for each patch, which were subsequently used as labels for machine learning. The role of machine learning in this context was to provide explainability and interpretability for the predictions made by the deep neural network. By analyzing the machine learning models through features, we aimed to gain insights into the underlying factors and patterns contributing to the EGFR mutation detection task. This is depicted in figure 3.

The research followed a systematic approach that involved several key steps shown in figure 6. These steps are included:

1. Generation of EGFR mutation Prediction Scores using DL model
2. Evaluation of EGFR mutation predictions using ML model

3.1. Dataset

This study utilized anonymized scanned Whole Slide Images (WSIs) obtained from the Genomic Data Commons (GDC) Portal [2], specifically retrieved from the The Cancer Genome Atlas (TCGA) Lung Adenocarcinoma (LUAD). The GDC Data Portal offers access to various types of TCGA data, including genomic, transcriptomic, proteomic, and clinical data derived from numerous tumor and normal tissue samples across multiple cancer types. Data retrieval can be performed through the GDC Data Transfer Tool or accessed via the GDC Data Portal API. The GDC Data Portal website provides comprehensive information about available data and instructions for data access and download, all of which adhered to the guidelines stipulated in the GDC Data Use Policy.

This study specifically focused on non-small cell lung cancer (NSCLC) with the TCGA LUAD dataset, with a specific emphasis on patients exhibiting EGFR mutations. The study comprised a total of 541 .SVS WSI slides, involving 522 patients (some patients having multiple slides), with 77 slides labeled as positive and 464 slides labeled as negative as shown in the figure 4.

[2] `https://portal.gdc.cancer.gov`

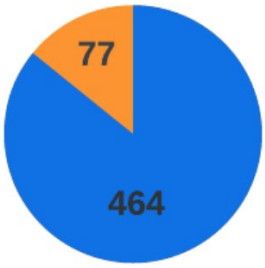


Figure 4: Distribution of whole-slide images per class, with the negative class represented by the color blue and the positive class represented by the color yellow

Experiments were run with a NVIDIA RTX A4000 graphic card and the following libraries : TensorFlow v2.8.0-rc0, keras v2.8.0, CUDA 11.5. The system specifications used for our machine learning models are displayed in table 1. Coding was done using Jupyter Notebook.

Table 1: System Specifications for running machine learning model

Specification	**Value**
Architecture	x86_64
CPU(s)	12
Model name	AMD Ryzen 5 3600 6-Core Processor
CPU max MHz	3600.0000

3.2. Generation of EGFR mutation prediction Scores using DL model

The initial phase of the research encompassed a series of sequential steps, which are outlined as follows:

1. Molecular Label Determination
2. Slide Preprocessing
3. Feature Learning
4. EGFR mutation prediction from histopathology

3.2.1. Molecular Label Determination

The molecular labels were determined based on the masked somatic mutations maf file, utilizing the MuTect236 algorithm specifically designed for the dataset. A gene was considered positive if it contained a mutation with an IMPACT value categorized as ”HIGH” or ”MODERATE” by the VEP software 37. Conversely, genes with other IMPACT values or no mutations were classified as negative. Positive labels were represented as 1, while negative labels were represented as 0. It is worth noting that there were no missing labels in the dataset.

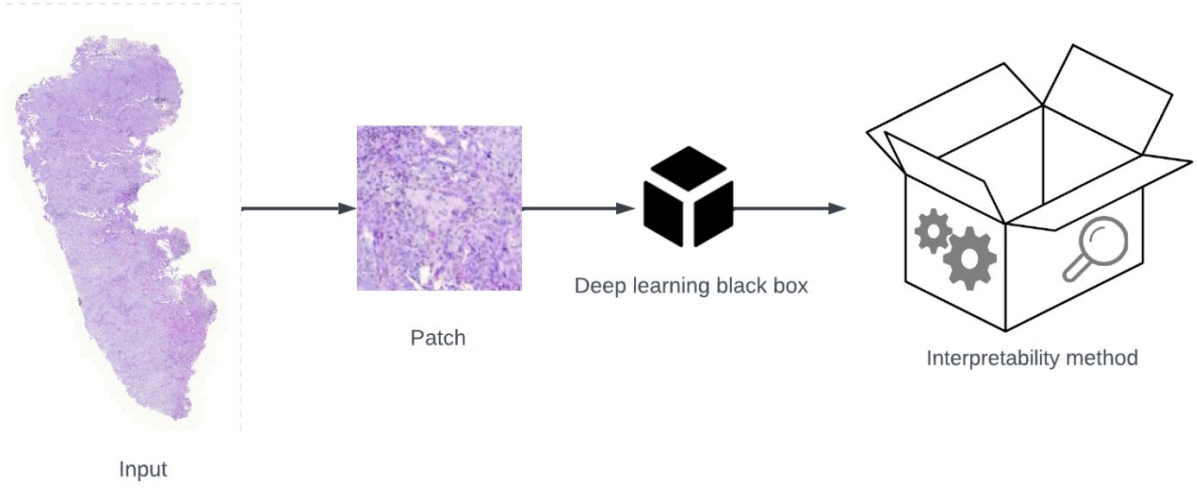


Figure 3: Overview of the strategy employed in this study for EGFR mutation detection. Whole slide images (WSIs) are divided into patches and fed into a deep neural network for prediction. The prediction scores from the deep neural network are used as labels for machine learning models, which provide explainability and interpretability. The machine learning models offer insights into the factors influencing EGFR mutation detection.

3.2.2. Slide Prepocessing

- Selection of Diagnostic Slides: The initial step involved selecting Aperio SVS files of formalin-fixed paraffin-embedded (FFPE) diagnostic slides from the dataset. These slides were identified by the presence of a ”DX” label in their filenames. No diagnostic slides were excluded from the subsequent analysis.

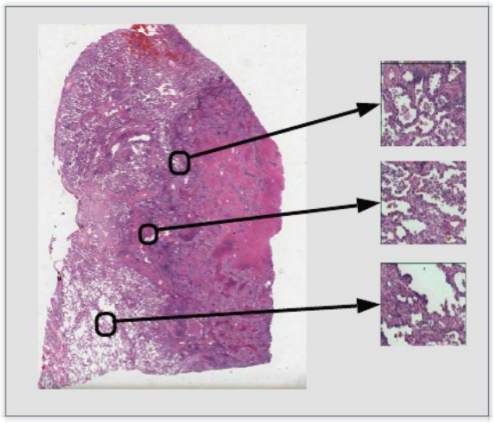


Figure 5: Non-overlapping patches are extracted from H&E (Hematoxylin and Eosin) Whole Slide Images (WSI) measuring 600*600 pixels at resolution ’*1’

- Foreground Extraction and Image Tiling: To prepare the images for further analysis, foreground extraction was performed using a U-net model trained in-house. Subsequently, the images were divided into non-overlapping patches measuring 600×600 pixels at a $1\times$ resolution. This is shown in figure 5. This process resulted in a total of 302,544 patches specifically for the TCGA-LUAD dataset. Each patch was associated with a label of 1 if the corresponding gene was mutated, and 0, if the gene was non-mutated, as determined by the ”Molecular Labels” section, described earlier.

- Standardization of Intensity Optical or Brightness: In order to normalize the light intensity across the images, centering and reducing all images to the average value for each RGB channel and their standard deviation is applied.

3.2.3. Feature Learning

We employed feature learning, which aims to discover meaningful representations within the data for effective categorization into desired classes. It automatically extracts features from images and condense them into a representation vector, also known as an embedding. In this study, an EfficientNetB7 neural network

was utilized, truncated at its last layer. A global average pooling layer was added to generate an embedding vector of size 2560.

3.2.4. EGFR mutation prediction from histopathology

A multilayer perceptron (MLP) was constructed to get the EGFR mutation prediction scores. The MLP consisted of two intermediate layers with 64 and 16 neurons, respectively, employing the rectified linear unit (ReLU) activation function. The final layer contained a single output neuron with a sigmoid activation function. It was trained for 5 epochs on batch size of 8 using the Adam optimizer with a learning rate of 1e-4 and the binary cross-entropy loss function. Subsequently, the trained model was applied to the test set, and the prediction values were averaged for each tile. Patch-level predictions were computed for each tile, and the process was repeated 5 times. The resulting predictions were used to calculate the area under the receiver operating characteristic curve (AUC).

3.3. Evaluation of EGFR mutation predictions using ML model

In this phase, the prediction scores generated by the deep learning model were employed as labels for the machine learning models. These ML models were identified through a literature review.

Initially, Before advancing to our core topic of research, we implemented a basic linear classifier using Pytorch in combination with in-house developed UDE library. This library is developed to facilitate data loading, structuring, pre and post processing on a standard procedure. Baseline code provided resourceful insights and knowledge necessary to complete the research work which would be discussed later in this chapter. This was performed on a set of 10 slides. Baseline linear classifier was designed to take 3 inputs in the form of mean RGB channel values and do binary classification.

As mentioned earlier this task proved to be very useful in integrating our code with the existing coding structure followed at the company which included working on patches generated from slide images, assigning proper labels using the UDE functions and get evaluation results. Table 4 summarizes the results of baseline program.

After working on baseline we performed following tasks on the regression model later on we tried them on classification model:

1. Data Preprocessing
2. Feature extraction
3. Data post processing
4. Machine learning

3.3.1. Data Preprocessing

Due to computational limitations, a subset of eighty slides was chosen from the initial 541 slides for further analysis. The subsequent steps involved both regression and classification tasks. The data preprocessing steps for the machine learning models were identical to those employed for the deep learning model.

- Regression Task using Prediction Scores: The initial phase of the study involved implementing a regression models. For this, the prediction scores obtained from the deep learning model were directly used as the target variable for the machine learning models.

- Visualization and Threshold Selection: Subsequently, the prediction scores were examined to determine slides with highest precision scores. This analysis facilitated the identification of an appropriate threshold for classifying the prediction scores into two distinct classes: 0 and 1. A total of 80 slides, comprising 40 negative and 40 positive slides, were selected as the dataset for further analysis. Here 0.4 was set as a threshold to convert the prediction score to binary labels. The results of the visualization process are depicted in the accompanying graph in 7.

  Utilizing the selected threshold, the prediction scores were binarized into two classes: 0 and 1. This enabled the classification task to proceed based on the prediction scores obtained from the machine learning models.

3.3.2. Feature Extraction

In accordance with the literature review and state-of-the-art (SOTA) approaches, feature extraction was performed on the RGB, HSV, and LAB color channels. The LAB and HSV color spaces were specifically utilized to enhance the representation of colors(Wang et al., 2017). The study commenced by extracting a set of 144 handcrafted features encompassing pixel intensity, texture, and color moments. Additionally, 236 Gabor features and 728 HOG features were incorporated, resulting in a total of 1208 features. Consequently, the experiments were conducted using two distinct feature sets: one comprising 144 features and the other consisting of 1208 features.

1. Texture Feature Extraction:
   It plays a significant role in cancer diagnosis within the field of radiomics, as they are closely associated with variations in protein expressions (Lambin et al., 2012). In this study, three distinct methods were employed for texture feature extraction: Gray-Level Co-Occurrence Matrix (GLCM), Local Binary Patterns (LBP), and Gabor filtering.
   Gray-Level Co-Occurrence Matrix (GLCM):

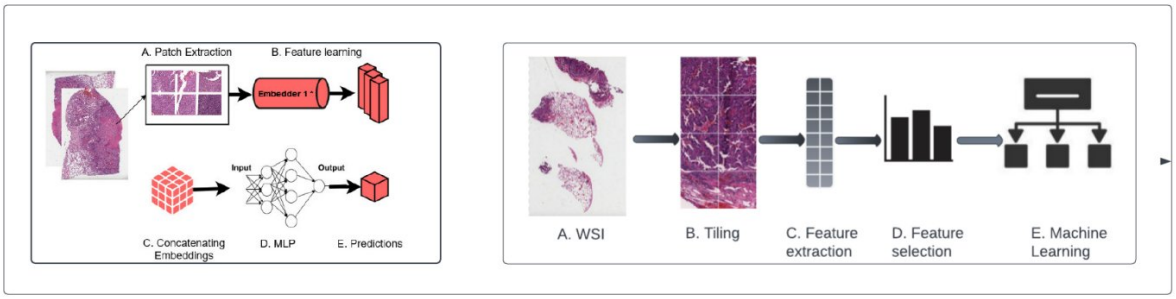


Figure 6: Overview of the applied strategy divided into two parts: **Generation of Prediction Scores (Left image): A**. Patches are extracted from the WSI measuring 600*600 pixels at resolution '*1'. **B).** EfficientNetB7 neural network was utilized, truncated at its last layer as feature extractor to generate an embedding vector of size 2560. **C).** Concatenate embeddings. **D).** Feed the embeddings to multi layer perceptron (MLP). MLP consisted of two intermediate layers with 64 and 16 neurons, respectively, employing the rectified linear unit (ReLU) activation function. The final layer contained a single output neuron with a sigmoid activation function. **E).** Prediction score of the deep learning model. **Machine learning prediction analysis (Right image): A).** H(Hematoxylin and Eosin)Whole Slide Images (WSI) from TCGA LUAD dataset. **B).** Patches created from tiling. **C).** Hand crafted features extracted including texture, color, pixel intensity and shape features. **D).** Random forest feature selection. **E).** Machine learning classification and regression models applied for predicting labels, further evaluated on R-squared score and AUC score.

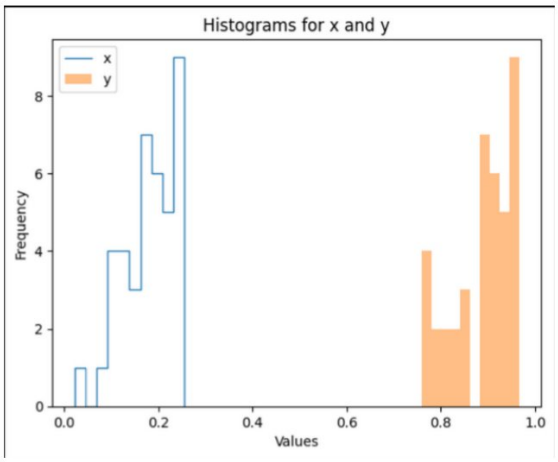


Figure 7: Visualization of prediction scores, x being labeled as 'class 0' (no EGFR mutation) and y (EGFR mutation) labeled as 'class 1'

GLCM relies on the spatial dependency of pixels within the gray-level co-occurrence matrix (Haralick et al., 1973). This matrix calculates the frequency of pixel pairs in the spatial domain and extracts statistical values. For this study, the GLCM features of contrast, correlation, homogeneity, and dissimilarity were specifically focused on, resulting in the extraction of 36 features from GLCM.

Local Binary Patterns (LBP):

LBP involves thresholding neighboring pixels and representing the result as a binary number (Ojala et al., 1996). It calculates the histogram of each cell within the neighboring pixels, combining them together, and performing normalization. In this study, LBP yielded 90 bins of features.

Gabor Filtering:

Gabor features are derived from Gabor filters, which are linear filters commonly used for edge detection. These filters excel at capturing spatial frequency information within images and are capable of incorporating details about the phase of spatial frequencies, which can be valuable for certain types of texture analysis (Manjunath and Ma, 1996) (Kruizinga and Petkov, 1999).

2. Color Moments:
   Color moments were extracted using a function that identified color markers through different thresholds applied to the image. Specifically, thresholds were set for colors such as black, red, blue, gray, white, light brown, and dark brown. This allowed for the capture of distinct color characteristics within the patches (Zhang and Lu, 2002).
3. Pixel Intensity Features:
   Pixel intensity statistics features were utilized to capture the color information of the patches. These features were computed based on the intensity values of the pixels within each patch. The calculated statistics included intensity mean, standard deviation, skewness, kurtosis, entropy, and energy. These measures provided insights into the distribution and variability of pixel intensities, contributing to the characterization of color properties (Zhao et al., 2018).
4. Shape Features:
   Histogram of Oriented Gradients (HOG):
   The Histogram of Oriented Gradients (HOG) represents the distribution of intensity gradients or edge directions within an image. The image is divided into smaller connected regions called cells. For each pixel within a cell, a histogram of gradient directions is constructed. These histograms are then normalized to enhance accuracy and improve the representation of shape-related informa-

tion (Dalal and Triggs, 2005).

3.3.3. Post-Processing

After the feature extraction, the data underwent post-processing steps including feature normalization and scaling to ensure consistent ranges and avoid bias towards certain features. Furthermore, a feature selection method was employed on both sets 144 and 1208, respectively to identify the most relevant features for the subsequent analysis. In this study, a random forest algorithm was utilized as the feature selector to identify the optimal set of features.

To confirm the robustness of the method, while avoiding overfitting on the same patient sets, we evaluated the strategies in all patients using repeated 5-fold cross-validation, configuration-I, and repeated the split-train-test (80:20 ratio), configuration-II. Configuration-I and II was applied for regression while configuration-II was only applied for classification tasks as illustrated in figure 8.

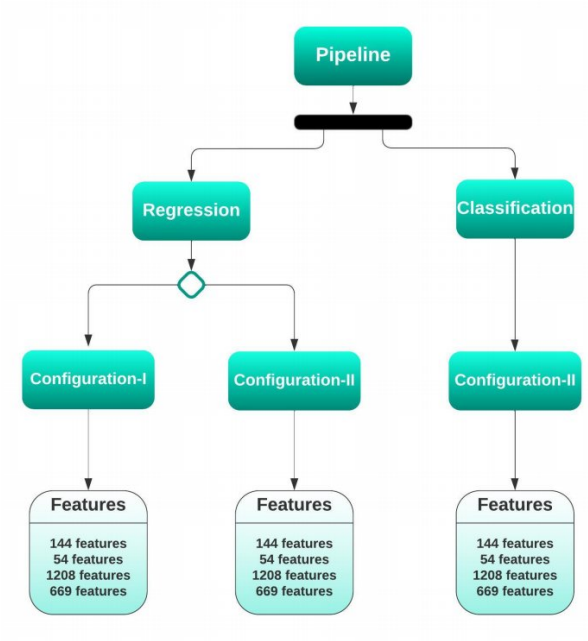


Figure 8: Evaluation pipeline illustrating the assessment of strategies using repeated 5-fold cross-validation (Configuration-I) and repeated split-train-test (80:20 ratio) approach (Configuration-II). Configuration-I was employed for regression tasks, while Configuration-II was utilized for both regression and classification tasks.

3.3.4. Machine Learning

To gain deeper insights and enhance the interpretability of the predictions, both regression and classification models were applied. To add on, initially regression and then classification models were applied. The following models were employed in this study:

- Support Vector Machine (SVM):

  SVM is a supervised learning algorithm that aims to find an optimal hyperplane to separate the labeled training data into different categories. It can effectively categorize new, untrained data based on the learned hyperplane.

  In our study, we experimented with SVM regression using both linear and radial basis function (RBF) kernels. We explored various C values ranging from 0.5 to 100 and observed that a value of 10 yielded the best results, as higher values led to overfitting. However, when it came to classification tasks, we excluded the linear kernel due to two main reasons. Firstly, the computation time required for the linear kernel was significantly high. Secondly, its performance in classification tasks was not satisfactory, prompting us to focus solely on the RBF kernel.

- Random Forest (RF):

  RF is a classification technique that employs multiple decision trees. Each decision tree is trained individually and then used in a randomized manner to classify untrained data. By combining the predictions of multiple decision trees, RF can provide robust classification results.

- XGBoost:

  XGBoost, short for Extreme Gradient Boosting, is a machine learning algorithm that utilizes the principle of boosting. It creates an ensemble of weak models that individually may not be strong predictors but, when combined, provide a powerful predictive capacity. XGBoost is particularly known for its efficiency and effectiveness in handling large datasets and complex problems.

3.4. Evaluation Metrics

The evaluation of the deep learning (DL) model was performed using the area under the receiver operating characteristic (AUC) curve as it provides a measure of the model's ability to discriminate between positive and negative instances.

For the machine learning (ML) regression models, the evaluation metric employed was the R-squared score. R-squared, also known as the coefficient of determination, it serves as an indicator of how well the model fits the data and captures the relationship between the predictors and the target variable.

Furthermore, in the case of the classifier models, the performance was assessed using the AUC curve, which plots the true positive rate against the false positive rate at different classification thresholds.

4. Results

Here we are presenting the compiled results of the strategy we employed in section 3: the study was done

Descriptors	Filters	no of filters	Feature names
Texture features	GLCM	36	contrast, dissimilarity, homogeneity, correlation
	Gabor	236	None
	LBP	90	None
Shape features	HOG	728	None
Pixel intensity features	None	108	mean intensity, standard deviation, skew, kurtosis, energy, entropy
Color Moments	None	12	black, red, blue gray, white, light brown, dark brown

Table 2: Features used for the analysis of predictions

in two parts, getting predictions from deep learning model and applying ML models. The research initially began with a baseline strategy, which served as the starting point for further exploration and improvement. Through iterative refinement, more sophisticated and refined strategies were developed and applied, leading to enhanced performance and insights in the research study.

4.1. EGFR prediction result using DL model

With reference to section 3, we tested the deep learning model for which the results depicted in table 3:

Table 3: Patch level AUC score for each fold of the DL model

Fold No.	1	2	3	4	5
AUC	0.68	0.60	0.78	0.52	0.84

The graph 9 illustrates the relationship between the area under the curve (AUC) and the number of epochs in the training process. Among the epochs tested, epoch five achieved the highest AUC of 0.85. On average, the deep learning model attained an AUC of 0.68 across the folds , demonstrating its capability to accurately classify the data. These results highlight the effectiveness of the deep learning model in learning and identifying patterns associated with the target variable.

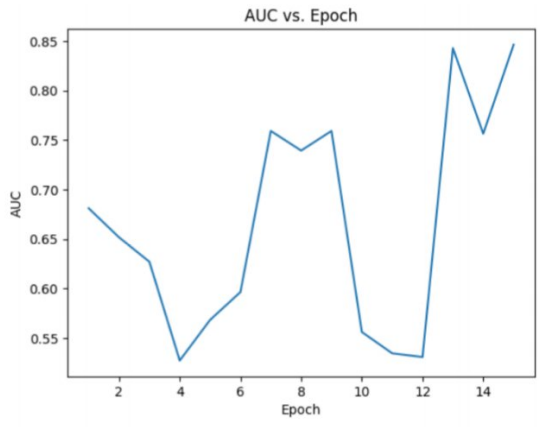


Figure 9: Patch level AUC vs Epoch graph

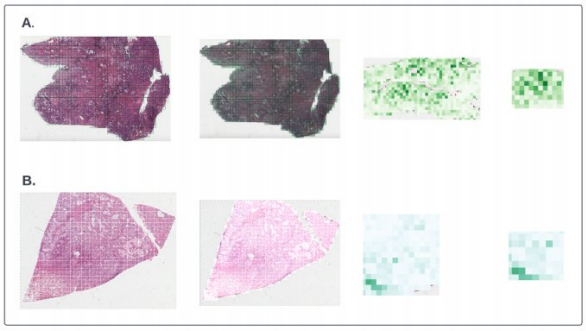


Figure 10: Left to right: A). Label visualization of a positively labeled Whole Slide Image (WSI) in green. Subsequently, the patch labels for the positively labeled slide are displayed, providing an insight into the distribution and classification of patches within the slide. B). Label visualization of a negatively labeled Whole Slide Image (WSI) in white. Subsequently, the patch labels for the negatively labeled slide are displayed, offering an understanding of the distribution and classification of patches within the slide.

As part of the results analysis, we utilized the Slide viewer of UDE to examine the visualizations of Whole Slide Images (WSI). This allowed us to gain insights into the distribution and characteristics of positive and negative patches within each slide. By analyzing the visualizations of both positive and negative WSIs, we gained insights into the discriminative features and regions that influenced the prediction results. The utilization of visual exploration techniques allowed for the observation of colors and intensities, even though not all aspects were clearly discernible as this requires the expertise of physicians. Nonetheless, this visual exploration served as a valuable complement to the quantitative analysis, facilitating a comprehensive interpretation of the model's performance shown in figure 10.

4.2. Baseline model results

The results presented in section 4 show the performance of the model when using the mean RGB values as input features. The evaluation was conducted on a dataset consisting of 10 whole slide images, with a total of 2,255 patches. The data was split into 80% for training and 20% for testing.

The model was optimized using the binary cross-entropy (BCE) loss function. The BCE loss achieved a value of 0.77, indicating the average discrepancy be-

tween the predicted and actual labels. The obtained average AUC value was 0.44.

These results suggest that using the mean RGB values as input features alone may not be sufficient for achieving high predictive accuracy. Further improvements and feature engineering techniques will be necessary to enhance the model's performance.

Table 4: Results of baseline linear classifier using Pytorch

Total Slides (Patches)	10 (2255)
train-test split	80:20
Batch Size	8
BCE Loss	0.77
Average AUC	0.44

4.3. EGFR mutation prediction evaluation using ML model

According to section 3 the second part of our strategy has been divided into two further parts: classification and regression. Initially, the research began with regression analysis but later transitioned to classification tasks. This shift was made to demonstrate the proof of concept and simplify the analysis by focusing on binary classification. The following section explains the results obtained from the regression experiments.

4.3.1. Regression result

The regression results obtained for both configuration-I and configuration-II are as follows:

For configuration-I figure 11, which involved the application of 5-fold cross-validation, the SVR model with a linear kernel and a C value of 10 outperformed the other models. The feature set used in this configuration consisted of 54 features, which were categorized into three groups: group 1 comprised 25 descriptors related to pixel intensity statistics, group 2 included 6 features reflecting color characteristics, and group 3 consisted of 23 texture features derived from gray-level co-occurrence texture matrices. The R-squared scores obtained for this model were 0.45, 0.22, -0.12, 0.1, and 0.01 for 5 CV folds, with an average of 0.132.

In configuration-II shown in figure 12, when applying the 80:20 split, the Random Forest model demonstrated superior performance compared to the other models. This model exhibited higher R-squared scores for both the 144 features and the selected 54 features. The feature sets used in this configuration were the same as in configuration-I. The Random Forest model achieved an R-squared score of 0.117 for the 144 features and 0.115 for the selected 54 features. These results indicate that the Random Forest model provided better insights into the explanations of the deep learning predictions and the features GLCM, LBP, pixel intensity and color moments are explaining the predictions better.

4.3.2. Classification result

ROC curve analysis results are shown in Figure 13 for the test set after applying the configuration-II. For the selected features, the AUC of RF machine learning method was high, with a maximum range of 0.667 for the test set having 144 features including GLCM, LBP, pixel intensity and color moments. These 144 features influenced the prediction model giving the AUC scores. Table 4.3.2 summarises the AUC for the machine learning models with different pipelines.

Table 5: AUC score for the 3 classifiers SVM= rbf, RF and XGboost

Pipeline	SVM	Random forest	XGBoost
144 features	0.6329	0.667	0.646
54 features	0.6329	0.662	0.646
1208 features	0.554	0.6102	0.601
669 features	0.498	0.6017	0.6124

Overall, the findings suggest that the Random Forest model performed well in explaining the deep learning predictions, particularly when utilizing a subset of informative features including GLCM, LBP, pixel intensity and color moments.

5. Discussion

The utilization of Whole Slide Images (WSI) in cancer research has emerged as a powerful tool for obtaining comprehensive and detailed information about tumor tissue. In this study, the researchers leverage the potential of WSI to enhance the understanding and interpretability of Epidermal Growth Factor Receptor (EGFR) mutations in lung adenocarcinoma (LUAD) samples through advanced machine learning (ML) techniques. The primary objective of this research is to develop a novel approach for inferring EGFR mutations using AI explainability and an ML model trained on the TCGA (The Cancer Genome Atlas) LUAD dataset. By employing a convolutional neural network (CNN) based architecture, the ML model learns to identify and recognize patterns associated with EGFR mutations within the WSI. The application of advanced ML techniques enhances the interpretability of EGFR mutations, providing insights into the underlying mechanisms and characteristics associated with LUAD.

In this research study, the performance of different machine learning models was evaluated using two approaches: 5-fold cross-validation and an 80:20 split. In terms of classification the results revealed that the linear Support Vector Regression (SVR) model performed better in the 5-fold cross-validation approach, while the Random Forest (RF) model outperformed other models in the 80:20 split approach. The SVR model achieved superior results with an average R-squared score of 0.132 on the 5 folds. On the other hand, the RF model demonstrated higher R-squared values with both 144

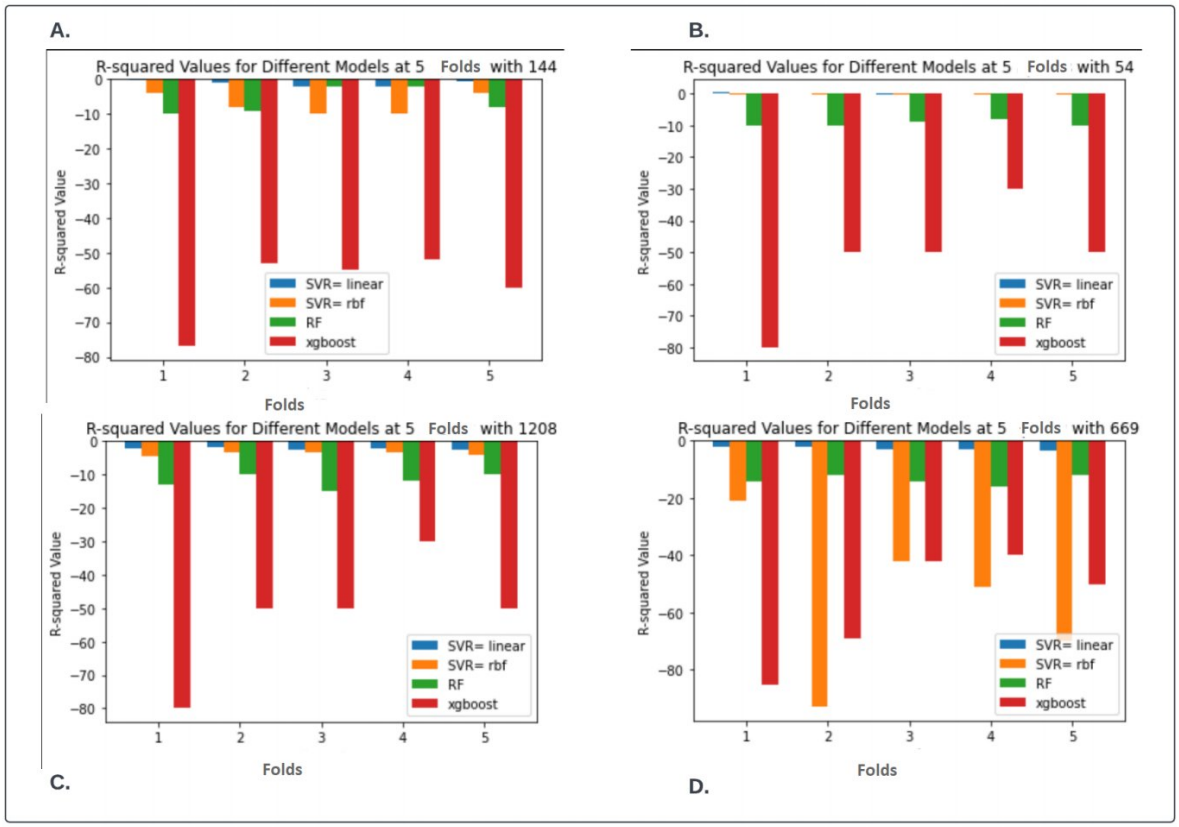


Figure 11: Comparison of R-squared scores for four models through 5-fold CV(SVR=linear, SVR=rbf, RF, and XGBoost) across different feature sets. The feature sets include A) 144 features, B) 54 features, C) 1208 features, and D) 669 features

features and the selected 54 features. The feature groups in this case were also pixel intensity statistics, color features, and texture features. In terms of classification results, the authors employed ROC curve analysis and reported the results. The RF model exhibited a higher AUC (Area Under the Curve) value, reaching a maximum of 0.667 for the test set when using the full set of 144 features. This indicates that the RF model had better discriminative ability in distinguishing between the positive and negative classes, explaining the deep learning model better. The selection of appropriate features from different groups (pixel intensity statistics, color features, and texture features) contributed to the improved performance of the models. It is worth noting that the performance of the RF model was consistently higher across both the regression (R-squared) and classification (AUC) analyses. In conclusion, the research demonstrated the potential of machine learning models, specifically SVR and RF, in accurately predicting and classifying EGFR mutations in LUAD samples using Whole Slide Images (WSI). By carefully selecting features from different groups, including pixel intensity statistics, color features, and texture features, the models were able to capture important patterns and relationships within the data, leading to improved understanding and interpretability of the predictions made by the deep learning model.

The results of this study demonstrate the potential of machine learning models, particularly SVR and RF, in accurately predicting and classifying the target variable. The combination of carefully selected features from different groups provides valuable insights into the underlying patterns and relationships within the data. Additionally, the analysis indicated that the set of 144 features and the selected subset of 54 features provided better explanatory power for the prediction results. These feature sets were able to capture important patterns and relationships in the data, leading to improved understanding and interpretability of the DL model's predictions.

The obtained results, although not optimal in terms of the R-squared score, can be attributed to several factors. Initially, the evaluation of our deep learning model was performed at the slide level, which may not have provided the best performance when applied to the patch level. It is important to note that a direct imitation of the deep learning model's performance is challenging, as its underlying convolutional operations differ significantly from traditional machine learning approaches. Moreover, the lack of available segmentation masks, com-

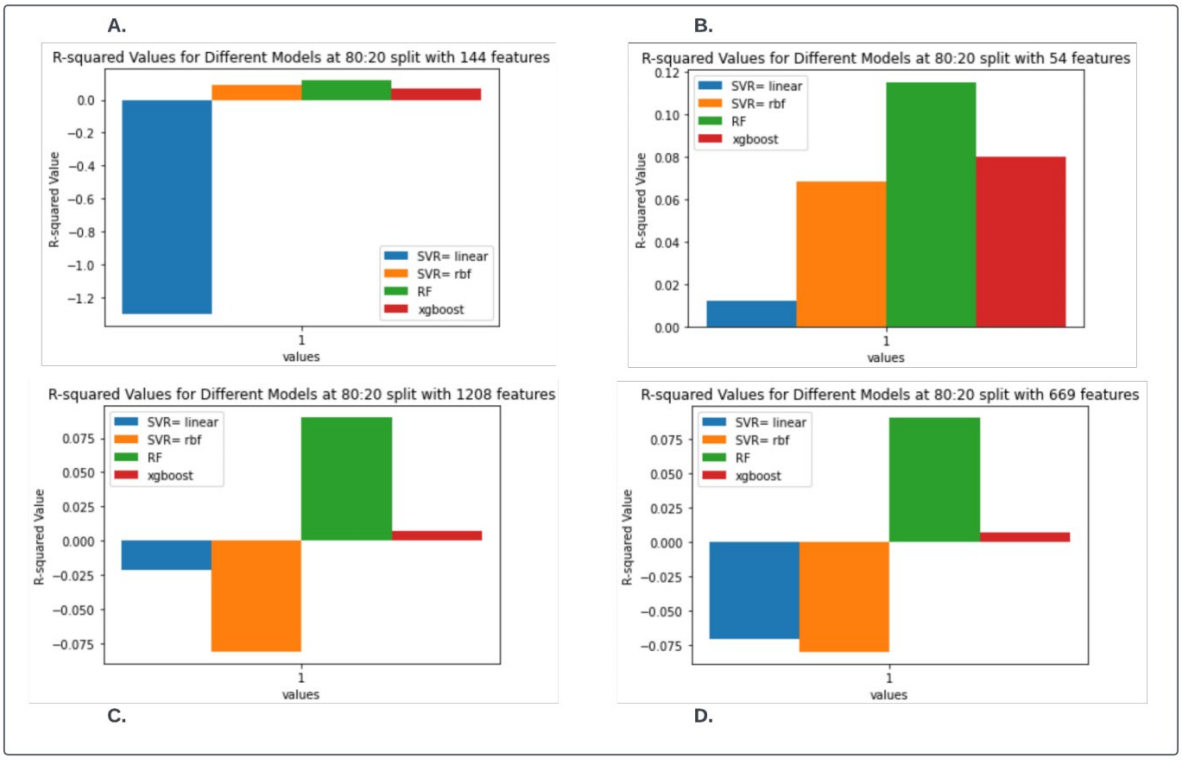


Figure 12: Comparison of R-squared scores for four models through 80:20 split (SVR=linear, SVR=rbf, RF, and XGBoost) across different feature sets. The feature sets include A) 144 features, B) 54 features, C) 1208 features, and D) 669 features

monly utilized in feature extraction according to the existing literature (Wang et al., 2017), resulted in a loss of information. These masks provide valuable insights for delineating regions of interest and extracting relevant features. The absence of such masks in our study may have limited the discriminative power of the extracted features. Given these limitations, further exploration and refinement of the methodology are warranted. Incorporating more accurate and precise annotations, such as expert-verified labels or improved segmentation masks, could enhance the reliability of the results. Additionally, investigating alternative approaches that leverage the strengths of both deep learning and traditional machine learning techniques may yield more robust models for this task.

6. Conclusions

This research aimed to develop a novel approach for inferring EGFR mutations in LUAD using AI explainability and ML models trained on the TCGA LUAD dataset. By employing a CNN-based architecture, the ML models learned patterns associated with EGFR mutations in WSI, enhancing interpretability. The study integrated handcrafted morphological features extracted from WSI images and applied regression and classification tasks. Future experiments should incorporate advanced radiomic and shape features near the tumor and employ supervised and unsupervised feature selection methods to capture complex relationships. Additional avenues for exploration include utilizing advanced deep learning techniques, combining imaging characteristics, and incorporating histology and pathology correlation analysis. Further studies can explore diverse strategies and employ various machine-learning techniques to optimize predictions using additional imaging features.

The results of the study demonstrated that the selection of appropriate features from different groups, including pixel intensity statistics, color features, and texture features, can contribute to the improved performance of the models in mutation prediction generated by deep learning. This suggests that these features provide a better explanation for the predictions made by the models.

However, it is important to acknowledge the limitations of the study. The evaluation of the deep learning model was performed at the slide level, which may not have provided optimal performance when applied to the patch level. The absence of available segmentation masks limited the discriminative power of the extracted features.

To address these limitations, future research should focus on incorporating more accurate annotations, such as expert-verified labels, and utilizing improved seg-

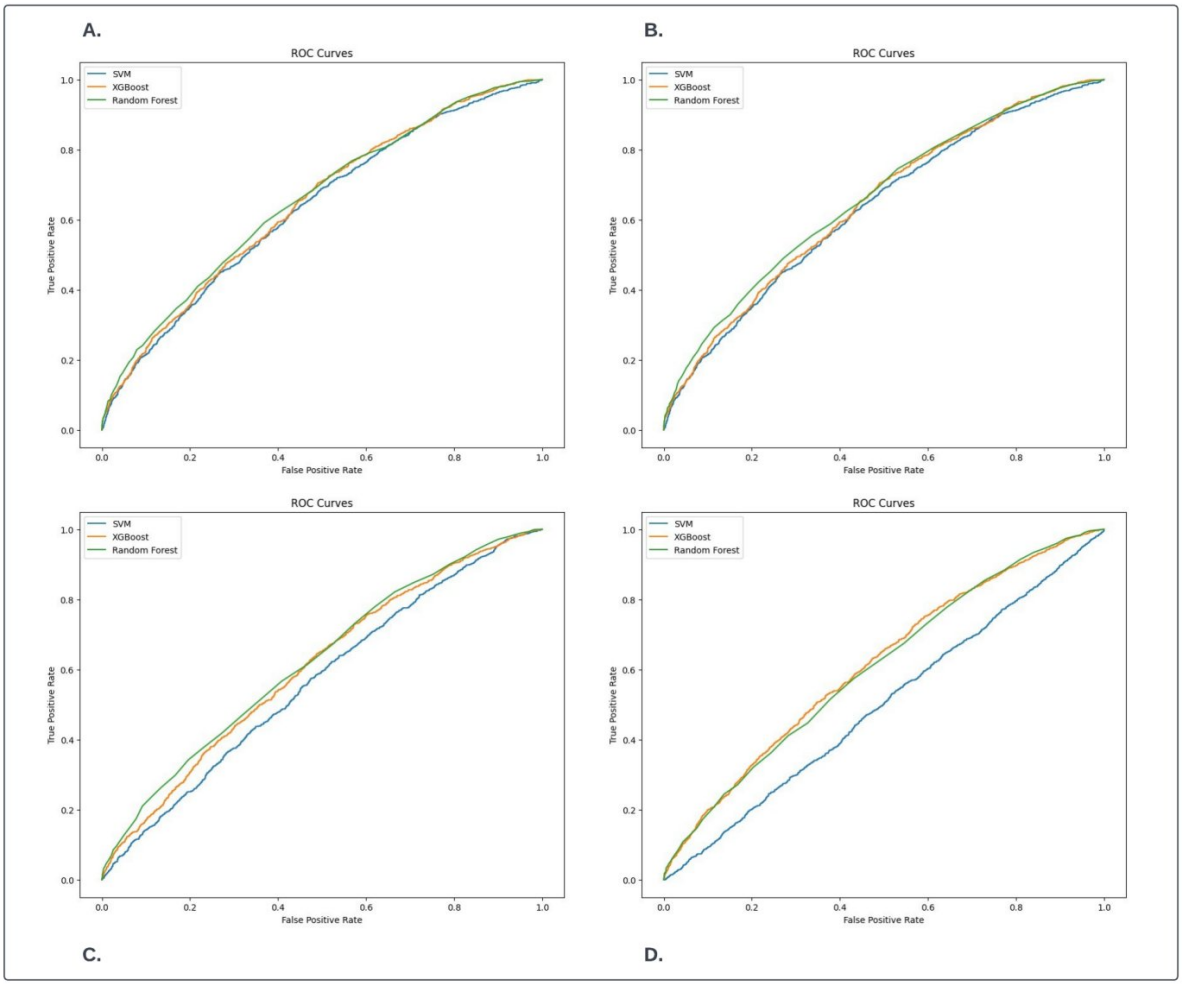


Figure 13: Comparison of AUC curves for three models through 80:20 split (SVC=rbf, RF, and XGBoost) across different feature sets. The feature sets include A) 144 features, B) 54 features, C) 1208 features, and D) 669 features

mentation masks to enhance the reliability of the results. Exploring alternative approaches that combine the strengths of deep learning and traditional machine learning techniques may also lead to more robust models for this task.

Acknowledgments

I would like to acknowledge and express my gratitude to God for providing me with the strength, guidance, and inspiration throughout this research journey.

I extend my sincere appreciation to my supervisor, Natalia, for her invaluable support, expertise, and guidance. Her insightful feedback and constructive suggestions have greatly contributed to the success of this project.

I would also like to acknowledge my co-supervisor, Nathan, for his assistance, encouragement, and valuable inputs. His expertise and knowledge have been instrumental in shaping this research.

I am grateful to Yann, my manager, for his support, encouragement, and the resources provided to carry out this research. His leadership and mentorship have been invaluable.

I would also like to express my gratitude to the entire research team for their collaboration, assistance, and valuable discussions.

Lastly, I would like to thank my family and friends for their unwavering support, understanding, and encouragement throughout this research endeavor. Their love and encouragement have been a constant source of motivation.

References

, . URL: `https://rb.gy/obg4k`.

Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S.S., Safwan, M., Alex, V., Marami, B., Prastawa, M., Chan, M., Donovan, M., et al., 2019. Bach: Grand challenge on breast cancer histology images. Medical image analysis 56, 122–139.

Arimura, H., Soufi, M., Kamezawa, H., Ninomiya, K., Yamada, M., 2019. Radiomics with artificial intelligence for precision medicine in radiation therapy. Journal of radiation research 60, 150–157.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion 58, 82–115.

Baik, C.S., Myall, N.J., Wakelee, H.A., 2017. Targeting braf-mutant non-small cell lung cancer: From molecular profiling to rationally designed therapy. Oncologist 22, 786–796.

Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B.E., Lee, B., Paeng, K., Zhong, A., et al., 2018. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. IEEE transactions on medical imaging 38, 550–560.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A., 2018. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians 68, 394–424.

Camidge, D.R., Kim, H.R., Ahn, M.J., Yang, J.C.H., Han, J.Y., Lee, J.S., Hochmair, M.J., Li, J.Y., Chang, G.C., Lee, K.H., Gridelli, C., Delmonte, A., Garcia Campelo, R., Kim, D.W., Bearz, A., Griesinger, F., Morabito, A., Felip, E., Califano, R., Ghosh, S., Spira, A., Gettinger, S.N., Tiseo, M., Gupta, N., Haney, J., Kerstein, D., Popat, S., 2018. Brigatinib versus crizotinib in alk-positive non-small-cell lung cancer. N Engl J Med 379, 2027–2039.

Campello, V.M., Gkontra, P., Izquierdo, C., Martin-Isla, C., Sojoudi, A., Full, P.M., Maier-Hein, K., Zhang, Y., He, Z., Ma, J., et al., 2021. Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. IEEE Transactions on Medical Imaging 40, 3543–3554.

Chan, B.A., Hughes, B.G., 2015. Targeted therapy for non-small cell lung cancer: current standards and the promise of the future. Translational Lung Cancer Research 4, 36–54. doi:10.3978/j.issn.2218-6751.2015.01.06.

Cireşan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J., 2013. Mitosis detection in breast cancer histology images with deep neural networks, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16, Springer. pp. 411–418.

Couraud, S., Zalcman, G., Milleron, B., Morin, F., Souquet, P.J., 2012. Lung cancer in never smokers–a review. European journal of cancer 48, 1299–1311.

Cruz, J.A., Wishart, D.S., 2006. Applications of machine learning in cancer prediction and prognosis. Cancer Informatics 2, 117693510600200030.

Cruz-Roa, A., Basavanhally, A., González, F., Gilmore, H., Feldman, M., Ganesan, S., Shih, N., Tomaszewski, J., Madabhushi, A., 2014. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks, in: Medical Imaging 2014: Digital Pathology, SPIE. p. 904103.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 886–893.

Digumarthy, S.R., Padole, A.M., Gullo, R.L., Sequist, L.V., Kalra, M.K., 2019. Can ct radiomic analysis in nsclc predict histology and egfr mutation status? Medicine 98.

Durán, J.M., Jongsma, K.R., 2021. Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical ai. Journal of Medical Ethics 47, 329–335.

Ettinger, D.S., Aisner, D.L., Wood, D.E., Akerley, W., Bauman, J., Chang, J.Y., Chirieac, L.R., D'Amico, T.A., Dilling, T.J., Dobelbower, M., Govindan, R., Gubens, M.A., Hennon, M., Horn, L., Lackner, R.P., Lanuti, M., Leal, T.A., Lilenbaum, R., Lin, J., Loo, B.W., Martins, R., Otterson, G.A., Patel, S.P., Reckamp, K., Riely, G.J., Schild, S.E., Shapiro, T.A., Stevenson, J., Swanson, S.J., Tauer, K., Yang, S.C., Gregory, K., Hughes, M., 2018. NCCN Guidelines Insights: Non-Small Cell Lung Cancer, Version 5.2018. Journal of the National Comprehensive Cancer Network 16, 807–821. doi:10.6004/jnccn.2018.0061.

Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D.M., Forman, D., Bray, F., 2015. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. International Journal of Cancer 136, E359–E386. doi:10.1002/ijc.29210.

Fu, Y., Jung, A.W., Torne, R.V., Gonzalez, S., Vöhringer, H., Shmatko, A., Yates, L.R., Jimenez-Linan, M., Moore, L., Gerstung, M., 2020. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nature cancer 1, 800–810.

Haralick, R.M., Shanmugam, K., Dinstein, I.H., 1973. Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics 3, 610–621.

Howlander, N., Noone, A., Krapcho, M., Miller, D., Brest, A., Yu, M., et al., 2019. Seer cancer statistics review (csr) 1975-2016. National Cancer Institute website seer cancer Published online .

Källén, H., Molin, J., Heyden, A., Lundström, C., Åström, K., 2016. Towards grading gleason score using generically trained deep convolutional neural networks, in: 2016 IEEE 13th International symposium on biomedical imaging (ISBI), IEEE. pp. 1163–1167.

Kather, J.N., Heij, L.R., Grabsch, H.I., Loeffler, C., Echle, A., Muti, H.S., Krause, J., Niehues, J.M., Sommer, K.A., Bankhead, P., et al., 2020. Pan-cancer image-based detection of clinically actionable genetic alterations. Nature cancer 1, 789–799.

Kruizinga, P., Petkov, N., 1999. Nonlinear operator for orientation and texture detection. IEEE Transactions on Image Processing 8, 457–467.

Lai, Y.H., Chen, W.N., Hsu, T.C., Lin, C., Tsao, Y., Wu, S., 2020. Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. Scientific reports 10, 4679.

Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., Van Stiphout, R.G., Granton, P., Zegers, C.M., Gillies, R., Boellard, R., Dekker, A., et al., 2012. Radiomics: extracting more information from medical images using advanced feature analysis. European journal of cancer 48, 441–446.

Levitsky, A., Pernemalm, M., Bernhardson, B.M., Forshed, J., Kölbeck, K., Olin, M., Henriksson, R., Lehtiö, J., Tishelman, C., Eriksson, L.E., 2019. Early symptoms and sensations as predictors of lung cancer: a machine learning multivariate model. Scientific Reports 9, 16504.

Lipton, Z.C., 2017. The doctor just won't accept that! arXiv preprint arXiv:1711.08037 .

Lipton, Z.C., 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16, 31–57.

Liu, Y., Kim, J., Balagurunathan, Y., Li, Q., Garcia, A.L., Stringfield, O., Ye, Z., Gillies, R.J., 2016. Radiomic features are associated with egfr mutation status in lung adenocarcinomas. Clinical lung cancer 17, 441–448.

Luo, X., Zang, X., Yang, L., Huang, J., Liang, F., Rodriguez-Canales, J., Wistuba, I.I., Gazdar, A., Xie, Y., Xiao, G., 2017. Comprehensive computational pathological image analysis predicts lung cancer prognosis. Journal of Thoracic Oncology 12, 501–509.

Manjunath, B., Ma, W.Y., 1996. Texture features for browsing and retrieval of image data. IEEE Transactions on Pattern Analysis and Machine Intelligence 18, 837–842.

Mei, D., Luo, Y., Wang, Y., Gong, J., 2018. Ct texture analysis of lung adenocarcinoma: can radiomic features be surrogate biomarkers for egfr mutation statuses. Cancer Imaging 18, 1–9.

Molina, J.R., Yang, P., Cassivi, S.D., Schild, S.E., Adjei, A.A., 2008. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship, in: Mayo clinic proceedings, Elsevier. pp. 584–594.

Ojala, T., Pietikäinen, M., Mäenpää, T., 1996. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns 1, 512–516.

Rossi, G., Barabino, E., Fedeli, A., Ficarra, G., Coco, S., Russo, A., Adamo, V., Buemi, F., Zullo, L., Dono, M., et al., 2021. Radiomic

detection of egfr mutations in nsclcradiomic detection of egfr mutations in nsclc. Cancer Research 81, 724–731.

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence 1, 206–215.

Schabath, M.B., Cote, M.L., 2019. Cancer progress and priorities: lung cancer. Cancer epidemiology, biomarkers & prevention 28, 1563–1579.

Shaw, A.T., Hsu, P.P., Awad, M.M., Engelman, J.A., 2013. Tyrosine kinase gene rearrangements in epithelial malignancies. Nat Rev Cancer 13, 772–787.

Shi, Y., Au, J.S.K., Thongprasert, S., Srinivasan, S., Tsai, C.M., Khoa, M.T., Heeroma, K., Itoh, Y., Cornelio, G., Yang, P.C., 2014. A prospective, molecular epidemiology study of egfr mutations in asian patients with advanced non-small-cell lung cancer of adenocarcinoma histology (pioneer). J Thorac Oncol 9, 154–162.

Sirinukunwattana, K., Domingo, E., Richman, S.D., Redmond, K.L., Blake, A., Verrill, C., Leedham, S.J., Chatzipli, A., Hardy, C., Whalley, C.M., et al., 2021. Image-based consensus molecular subtype (imcms) classification of colorectal cancer using deep learning. Gut 70, 544–554.

Song, Y., Zou, J.J., Chang, H., Cai, W., 2017. Adapting fisher vectors for histopathology image classification, in: 2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017), IEEE. pp. 600–603.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR. pp. 6105–6114.

Temme, M., 2017. Algorithms and transparency in view of the new general data protection regulation. European Data Protection Law Review 3. URL: https://doi.org/10.21552/edpl/2017/4/9.

Terra, S.B., Jang, J.S., Bi, L., Kipp, B.R., Jen, J., Eunhee, S.Y., Boland, J.M., 2016. Molecular characterization of pulmonary sarcomatoid carcinoma: analysis of 33 cases. Modern Pathology 29, 824–831.

Thawani, R., McLane, M., Beig, N., Ghose, S., Prasanna, P., Velcheti, V., Madabhushi, A., 2018. Radiomics and radiogenomics in lung cancer: a review for the clinician. Lung cancer 115, 34–41.

Travis, F., SecondAuthor, ThirdAuthor, FourthAuthor, 2015. Title of the article. Journal Name X, Y–Z.

Vargas, A.J., Harris, C.C., 2016. Biomarker development in the precision medicine era: lung cancer as a case study. Nature Reviews Cancer 16, 525. doi:10.1038/nrc.2016.56.

Wang, C., Su, H., Yang, L., Huang, K., 2017. Integrative analysis for lung adenocarcinoma predicts morphological features associated with genetic variations., in: PSB, World Scientific. pp. 82–93.

Yamashita, R., Long, J., Saleem, A., Rubin, D.L., Shen, J., 2021. Deep learning predicts postsurgical recurrence of hepatocellular carcinoma from digital histopathologic images. Scientific reports 11, 1–14.

Yang, J., Veeraraghavan, H., Armato III, S.G., Farahani, K., Kirby, J.S., Kalpathy-Kramer, J., van Elmpt, W., Dekker, A., Han, X., Feng, X., et al., 2018. Autosegmentation for thoracic radiation treatment planning: a grand challenge at aapm 2017. Medical physics 45, 4568–4581.

Yu, K.H., Zhang, C., Berry, G.J., Altman, R.B., Ré, C., Rubin, D.L., Snyder, M., 2016. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nature communications 7, 12474.

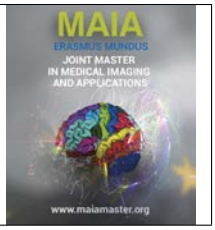
Zhang, D., Lu, G., 2002. Review of shape representation and description techniques. Pattern Recognition 35, 433–455.

Zhao, Z., Yang, Z., Jiang, Z., Luo, J., 2018. Combining texture and intensity features for texture classification. International Journal of Machine Learning and Cybernetics 9, 659–668.

Zhuang, X., Li, L., Payer, C., Štern, D., Urschler, M., Heinrich, M.P., Oster, J., Wang, C., Smedby, Ö., Bian, C., et al., 2019. Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. Medical image analysis 58, 101537.

Medical Imaging and Applications

Master Thesis, June 2023



OIDA - Optic disc segmentation with Image-to-image translation for Domain Adaptation

Christina Bornberg, Damon Wong, Jacqueline Chua, Leopold Schmetterer

Singapore Eye Research Institute

Abstract

Introduction: Understanding the vascular and neuronal components of the retinal nerve fibre layer around the optic disc is crucial in understanding eye diseases such as Glaucoma. Previous work (Yow et al., 2021) established a pipeline for optic disc segmentation, layer segmentation and vessel extraction using manual annotation, traditional image processing as well as deep learning steps. *Objective:* The goal of this work is to automate the optic disc segmentation. A manually annotated source-domain dataset is available, while the target-domain dataset is unlabelled.

Methods: The semi-supervised approach "Mean Teacher" was used in order to learn from both labelled and unlabelled data. *Results:* We implement three experiments, a baseline mean teacher pipeline, a mean teacher with Fourier domain adaptation image-to-image translation, and a third pipeline, using a mean teacher, Fourier domain adaptation, with an additional focus on uncertainty, where we introduce both, cleaning the teacher mask as well as using entropy of a test-time augmentation setup as a loss term. We achieve 90% f-score on our final pipeline. Additionally, our method has low computational cost, applying an Efficientnet B0, training for 5.6 hours.

Conclusion: Fourier domain adaptation (FDA) works decent accross all domains. For multi-modal unsupervised image-to-image translation (MUNIT), we need to do further investigation.

Keywords: Domain Adaptation, Semantic Segmentation, Optical Coherence Tomography

1. Introduction

1.1. Towards understanding the circumpapillary retinal nerve fibre layer

Glaucoma is a leading cause of irreversible blindness worldwide (Pascolini and Mariotti, 2012). It is a group of diseases characterised by the progressive loss of retinal ganglion cells (RGC) which causes changes in the optic nerve head (ONH) and retinal nerve fibre layer (RNFL) (Sharma et al., 2008).

Once the vision is compromised, it cannot be restored, but it is possible to control and prevent further deterioration of vision. If identified early, the disease progression can significantly be slowed down with medical and surgical therapy (Senjam, 2020). If left untreated, eventually, glaucoma leads to visual dysfunction and blindness.

Many people who have glaucoma are unaware of it because symptoms do not usually occur during the early stage of the disease (Chua et al., 2015). By the time, patients notice some signs and symptoms, the disease has already caused irreparable damage. It is estimated that about 70-75% of glaucoma patients are undiagnosed (Heijl et al.; Weih et al., 2001). Without some form of screening, most patients with glaucoma remain undiagnosed until an advanced disease stage is reached (Tan et al., 2020).

Unfortunately, the pathogenesis of glaucoma is not fully understood (Weinreb et al., 2014). Proposed risk factors are an increased intraocular pressure (IOP), aging, and family history, however, vision loss can occur with normal pressure or even lower. Reduction of intraocular pressure is the only proven method to treat the disease.

Guo et al. (2005) found that there is a relation between retinal ganglion cell death, intraocular pressure and IOP-induced effects on the extracellular matrix. This is confirmed by Weinreb et al. (2014) in their re-

view study, stating, that the level of intraocular pressure is related to retinal ganglion cell death.

Retinal ganglion cells die when their axons, that form the optic nerve, are injured (Sánchez-Migallón et al., 2016). Those axons of the retinal ganglion cells are the primary component of the retinal nerve fiber layer (RNFL). Due to their progressive degeneration in glaucoma, the layer becomes thinner. Additionally, vessels are present in this layer, that influence the thickness measurement (Yow et al., 2021).

In order to diagnose Glaucoma, the imaging technique optical coherence tomography (OCT) can be applied (Sharma et al., 2008). OCT devices are able to image the RNFL around the optic nerve head, also known as circumpapillary retinal nerve fiber layer (cpRNFL). However, OCT scans alone do not clearly distinguish between the neuronal and vascular components within the RNFL.

Fortunately, Optical Coherence Tomography Angiography (OCTA) (Spaide et al., 2018) was introduced, a non-invasive imaging modality building on OCT that is able to provide depth-resolved images of blood flow in the retina and choroid.

Previous work by Yow et al. (2021) focused on combining both OCT and OCTA image pairs in order to understand the vascular and nerve components of the cpRNFL. Their study relies on the OCTA scanner CIRRUS, a research imaging device by ZEISS.

In our current study, we aim to bridge this gap by translating the approach of segregating the vascular and nerve components of the RNFL from the prototype OCTA system PLEX ELITE to the commercially available OCTA system CIRRUS. By adapting the methodology to a widely used OCTA system, we can enhance its clinical applicability and enable healthcare professionals to obtain valuable information about glaucoma's vascular and neuronal aspects. This translation will facilitate the integration of this technique into routine glaucoma diagnostics, enabling more accurate and comprehensive assessments of the disease.

1.2. Current pipeline

This work is a follow-up project of "Segregation of neuronal-vascular components in a retinal nerve fiber layer for thickness measurement using OCT and OCT angiography", (Yow et al., 2021). Yow et al. (2021) focus on the understanding of the circumpapillary retinal nerve fibre layer (cpRNFL) of healthy eyes imaged by a ZEISS PLEX OCTA device.

In order to extract the thickness of the circumpapillary retinal nerve fibre layer (cpRNFL) while excluding vessels and only focusing on the neuronal components, a range of steps are applied. The current pipeline consists of a range of manual and/or computational steps.

First, bscans are extracted from both the OCT and OCTA volumes, meaning we can see the different layers of the eye.

In the next step, we generate an enface image by averaging each volume from the top view. The enface image shows the optic nerve head and vessels similar to a fundus image.

Now, a medical expert needs to segment the optic disc which is subsequently used to extract the centre point of the optic disc.

With the help of the centre point, a circumpapillary scan is extracted at a radius of 3.46 mm, this scan shows the layers similar to a bscan. Averaging over a range of radii increases the robustness of the scan.

In the same step, we get additional information from a superficial enface scan, which can be exported from the CIRRUS device. The vessel information from the same radius of 3.46 mm are stored in a vector. We refer to this 2D vector as *vessel information of the superficial image*.

Coming back to the circular scan, the next step is the manual layer segmentation of the retinal nerve fibre layer of the OCT image in order to extract the layer of interest. We refer to this 2D image as *layer mask*.

The layer mask is used to also extract the layer in the circular OCTA scan. By thresholding, we receive vessel information. We refer to this 2D image as *vessel information derived from the OCTA image*.

In the last step, we combine the *layer mask* with the *vessel information derived from the OCTA image* and the *vessel information of the superficial image*. From this, measurements of the layer, with and without vessels can be extracted.

1.3. Problem Statement

In this work, we will focus on the problem of finding the centre point of the optic disc. This task is currently done manually, which has the disadvantage that it is time-consuming. An alternative approach would be image processing, with the disadvantage that it is not robust against noise or illumination changes. A supervised deep learning approach would require a labelled dataset, which is not available. Labels in other domains are available, but come with a domain shift to our dataset.

1.4. Proposed Solution

We introduce a pipeline (see Figure 1), using common domain adaptation techniques, as well as semi-supervised learning for unsupervised domain adaptation. We have three datasets available, one being the target domain, and two different source domains. We transform the images in order to match the style of the target domain. We compare different approaches. A mixed batch sampling strategy makes sure that there is an equal amount of data from each dataset in each batch. A modified mean teacher approach is used for semi-supervised learning.

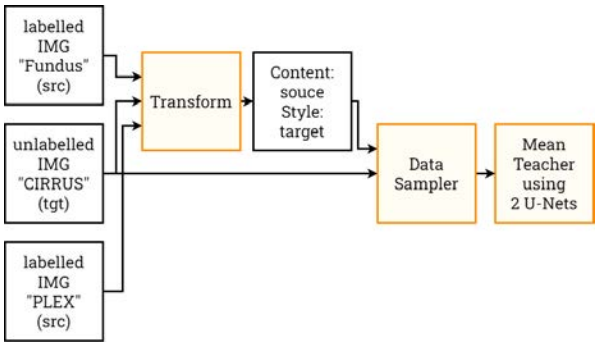


Figure 1: Segmentation pipeline using domain adaptation approaches combining labelled (source-domain) and unlabelled (target-domain) data. A transform is applied in order to reduce the domain shift. A sampling approach is applied in order to avoid an imbalance between the datasets. A Mean Teacher, consisting of a student and teacher model is used for temporal ensembling and pseudo label generation, hence learning from the unlabelled data.

1.5. Important concepts

For better understanding, the following terms are described.

Domain shift occurs if the distribution of images in the source dataset is different to target dataset (Liu et al., 2022).

Domain-specific representations describe the style of an image. Precisely, this could be colour, intensities, and noise. This information may be encoded as a vector (Liu et al., 2022).

Domain-invariant representations describe the content of an image, hence the geometry. In order to preserve spatial correlations, this can be encoded in a spatial spatial (tensor) (Liu et al., 2022).

Unsupervised domain adaptation is an umberella term for techniques that reduce the domain gap between a source and a target domain. This can be on input-level, on feature-level, or on output-level (Toldo et al., 2020).

Image-to-image (I2I) translation is a technique for UDA and has the goal to translate one image representation into another where a specific factor differs (e.g. style) while others are maintained (Liu et al., 2022).

Semi-supervised learning is a type of machine learning that combines both labelled and unlabelled data to improve model performance. Semi-supervised learning can be applied to unsupervised domain adaptation.

Semi-supervised learning for unsupervised domain adaptation. To bring both parts into context, unsupervised domain adaption can be considered as a variant of semi-supervised learning, but with a statistical shift of the unlabelled target data as additional complexity. Pseudo-labels implicitly promote feature-level cross-domain alignment, while still retaining the task specificity (Yang and Soatto, 2020). However, a common challenge in semi-supervised learning with pseudo-label generation is the tendency to prioritise more confident predictions. This becomes problematic when dealing with domain shifts as high confidence can be misleading due to the presence of out-of-distribution data with low certainty. To address this issue, pseudo-labelling approaches often employ techniques for cleaning the labels. This process may involve using uncertainty, for example measured with entropy, to ensure more reliable predictions (Toldo et al., 2020).

Semi-supervised domain adaptation (Toldo et al., 2020) is a combination of unsupervised domain adaptation and semi-supervised learning. Here, the source domain is fully labelled and the target domain is partially labelled. We do not apply this in our work, but want to mention it to avoid confusion due to the similar sounding terms in literature.

2. State of the art

We analyse approaches that address the domain shift and/or the unlabelled image problem for semantic segmentation. Figure 2 visually puts the following approaches into context that we distinguish between:

- Approaches that are designed for multi-modality or general robustness/domain invariance.
- Approaches that apply supervised or semi-supervised learning.
- Approaches that are used standalone/pre-trained or are learned.

Approaches may include generative and/or adversarial techniques.

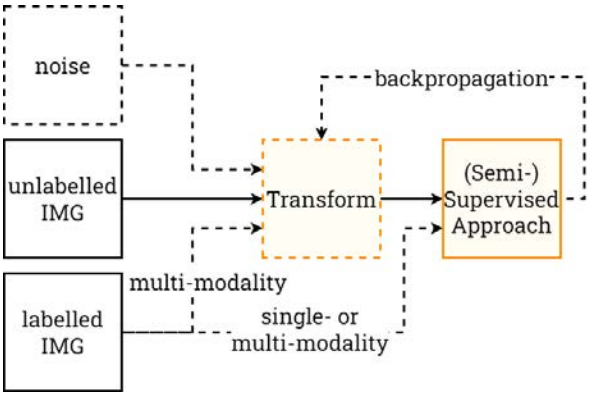


Figure 2: Overview of related work, including multi-modal or single-modal data, noise for robustness, transforms (for instance image-to-image translation or data augmentation), and learning approaches such as supervised or semi-supervised learning. We limit our related work to pipelines that combine labelled and unlabelled images. Dashed lines represent optional paths that vary between related work.

Generally said, an unlabelled image is transformed based on noise and/or labelled images to reduce the domain shift and then used for training a segmentation network. A combination of the transform and supervised learning may use the content of a labelled (source) image and the style of an unlabelled (target) image to produce synthecised labelled images in the target domain.

Changing the supervised to a semi-supervised learning approach, real unlabelled images in the target domain can be additionally introduced.

2.1. Transforms

Transforms can be either standalone, including pre-trained or trained in the final semi-(supervised) routine.

A rather simple approach is introducting noise for robustness.

Another approach that can be applied here carries the name unsupervised image-to-image translation. The goal here is to translate one image representation into another where a specific factor differs (e.g. style) while others are maintained.

2.1.1. Noise for adaptation robustness

Adding noise to make a domain adaptation pipeline more robust can be as simple as applying a Gaussian noise layer and can be as complex as applying adversarial data augmentation.

Volpi et al. (2018) propose an adversarial data augmentation technique.

Chen et al. (2022) developed a method for realistic adversarial data augmentation framework "AdvChain" for medical image segmentation tasks. Their method jointly optimises a dynamic data augmentation module and a segmentation network in order to better leverage labelled and unlabelled data. While not being designed for multi-modal data, as there is no knowledge about the target domain integrated, this work makes the model generally more robust.

2.1.2. Fourier domain adaptation

(Learnable) Fourier domain adaptation (FDA) (Yang and Soatto, 2020) is a simple type of unsupervised domain adaptation (UDA), using the Fourier domain for swapping the low-frequency spectrum between two domains.

2.1.3. Correlation alignment

CORrelation ALignment (CORAL) (Sun et al., 2017) is a simple method for unsupervised domain adaptation. Domain shifts are reduced by aligning the second-order statistics of source and target distributions.

2.1.4. Generative adversarial networks

Generative adversarial networks (GANs) (Goodfellow et al., 2020) can be applied to reduce the domain shift impact of a source and target domain. GANs typically employ a generator and a discriminator network. The generator generates an image by sampling from a Gaussian distribution, while the discriminator is given the synthetic image and a real one, and tries to identify which input is real and which is fake. Over the years multiple versions for image-to-image translation established, including Cycle-GAN (Zhu et al., 2017), Style and Content disentangled GAN (SC-GAN) (Kazemi et al., 2019), or the conditional generative adversarial network (cGAN) based approach pix2pix (Isola et al., 2017).

2.1.5. Unsupervised image-to-image translation network

Unsupervised image-to-image translation networks come in a range of versions. The basic version, the unsupervised image-to-image translation (UNIT) network (Liu et al., 2017) learns a one-to-one mapping between two visual domains. The multi-modal unsupervised image-to-image translation (MUNIT) network (Huang et al., 2018) learns a many-to-many mapping between two visual domains. The few-shot unsupervised image-to-image translation (FUNIT) network (Liu et al., 2019) learns a style-guided image translation model that can generate translations in unseen domains. The few-shot Unsupervised Image Translation with a content conditioned style encoder (COCO-FUNIT) network (Saito et al., 2020) is a FUNIT, introducing a content-conditioned style encoding scheme.

2.2. Semi-supervised learning

Over the past decades, there were many approaches introduced to tackle semi-supervised learning. Semi-supervised learning (SSL) is halfway between supervised and unsupervised learning, using both, labelled and unlabelled data for training.

We want to introduce three main concepts, entropy minimisation, consistency regularisation and pseudo labelling. As well as multiple architectures that use these strategies as a foundation.

Entropy minimisation (Grandvalet and Bengio, 2004) incorporates uncertainty additionally to confidence into the training routine to receive more certain predictions. In its simplest way is performed at pixel-level, so that each spatial unit of the prediction map brings an independent contribution to the final objective.

Consistency regularisation follows the assumption, that a perturbation, for instance, dropout, data augmentation or multiple models trained together, should not modify model predictions given the same input.

Pseudo-labelling methods use a model, trained on the labelled set to produce additional training examples by labelling images of the unlabelled set.

Self-training (which became popular in the 1970s through the success of the expectation maximisation algorithm), and later co-training (Blum and Mitchell, 1998) use pseudo labelling. Either a single or multiple networks generate pseudo labels for further training iterations.

Temporal ensembling (Laine and Aila, 2016) employs self-ensembling. This means that a prediction of unlabelled images are derived using the outputs of the

network-in-training on different epochs using different regularisation and data augmentation techniques.

The Mean Teacher (Tarvainen and Valpola, 2017) algorithm uses the strategy of temporal ensembling in combination with an exponential moving average between two models. This can improve consistency regularisation and reduces overfitting. One of its drawbacks is that given a large number of epochs, the teacher model's weights converge to those of the student model, the result is that biased and unstable predictions are carried over to the student.

Virtual Adversarial Training (VAT) (Miyato et al., 2018) is a semi-supervised learning method based on adversarial noise. Adversarial noise is injected into the training data for consistency regularisation. Using a confidence threshold, pseudo labels can be obtained. It can improve the generalisation performance and reduce the influence of noisy labels.

MixMatch (Berthelot et al., 2019) performs linear interpolation to mix both labelled and unlabeled images to get augmented image-label pairs.

FixMatch (Sohn et al., 2020) enforces the prediction consistency between weakly augmented images and strongly augmented images.

2.3. Domain-invariant learning

A very common approach to handle the domain adaptation problem is to apply adversarial learning. The theory of using adversarial learning for domain adaptation follows a simple approach: a model acts discriminative for the main learning task, on the source domain and indiscriminate with respect to the shift between the domains.

Ganin et al. (2016) introduced the domain-adversarial neural network (DANN). They combine a feature extractor and a label predictor to form a common feed-forward architecture. An additional domain classifier is connected to the feature extractor by a gradient reversal layer, enabling domain-invariant feature learning.

2.4. Semgnetation Pipelines

Zhang et al. (2020) developed a semi-supervised pipeline for unsupervised domain adaptation applying label propagation with augmented anchors.

Choi et al. (2019) introduced target-guided and cycle-free data augmentation (TGCF-DA), a GAN-based augmentation method for domain alignment. The final augmented image has the content of the labelled image, adapted to the style of the unlabelled image. In the second step, they use the labelled real data and augmented data as well as the unlabelled data in a semi-supervised semantic segmentation setup.

Wang et al. (2020) propose a fine-grained adversarial learning framework for cross-domain semantic segmentation. Their contribution is a "fine-grained" discriminator that can both, distinguish between domains and capture class knowledge in order to support feature alignment.

Chen et al. (2020) proposed a generative approach, where a synthetic dataset is used for training an image segmentation network. With the help of a MUNIT, they translate images from the labelled source domain into the unlabelled target domain. While their cascaded U-Net setup has never seen real labelled images before, it generalises well on the target domain.

Qin et al. (2023) proposed an unsupervised domain-adaptation pipeline for semantic segmentation. The approach makes use of the semi-supervised approach "dual student", as well as adversarial training.

Ouyang et al. (2019) proposed a pipeline that combines VAE-based feature prior matching with domain adversarial training. The goal is to learn a shared domain-invariant latent space which is then used for segmentation.

Zeng et al. (2021) proposed a pipeline that applies a CycleGAN for image translation and a domain-specific segmentation module.

Zhao et al. (2019) developed the Multi-source Adversarial Domain Aggregation Network (MADAN), a pipeline that combines dynamic adversarial image generation, adversarial domain aggregation, and feature-aligned task learning.

Yang and Soatto (2020) proposed a pipeline that combines a Fourier domain adaptation (FDA) module that reduces the domain gap between source and target without training required, and a segmentation network.

3. Material and methods

A general overview on how the datasets, transform, data sampler and mean teacher are connected is visually presented in Figure 1. As part of the mean teacher approach, the student model, trained through backpropagation, can be seen in Figure 7, and the teacher model, updated through the exponential moving average, can be seen in Figure 10.

3.1. Datasets and pre-processing

For the experiments, we use three different datasets. Zeiss PLEX Elite OCT enface scans as the labelled source domain, publicly available iChallenge datasets with labelled fundus images as helper domain, and finally, the Zeiss CIRRUS AngioPLEX OCT enface scans as the target domain.

The PLEX dataset (see Figure 3) consists of healthy eyes (270 images). Pre-processing steps include extracting the enface scan from the OCT volume. One volume has a size of 500x500x1536, covering the region of 6mm x 6mm. Binary masks with the optic disc as foreground class were provided by the lab. The validation set consists of 20% of the images.

The ADAM iChallenge (see Figure 4) dataset consists of AMD/non-AMD fundus images (Fu et al.,

2020). Pre-processing steps include cropping the fundus images around the optic disc according to the mask. Binary masks with the optic disc as foreground class are part of the dataset. Images with empty masks are excluded from training the pipeline, the final number of images used is 270.

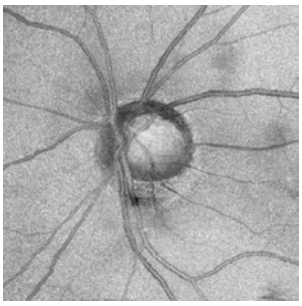


Figure 3: Example PLEX image.

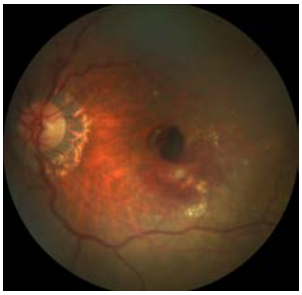


Figure 4: Example fundus image of the ADAM dataset.

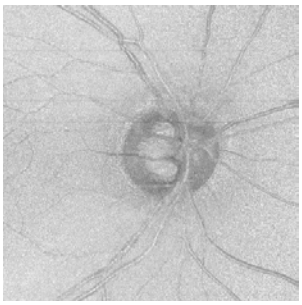


Figure 5: Example CIRRUS image.

The CIRRUS dataset (see Figure 5) consists of Glaucoma/non-Glaucoma eyes (1347 images). Pre-processing steps include extracting the enface scan from the OCT volume. One volume has a size of 350x350x1024, covering the region of 6mm x 6mm. No manually annotated masks are available for training and validation. A testset of 50 images were manually annotated.

3.2. Domain shift analysis

3.2.1. Visually

One way to understand the domain shift between source and target domain(s) is by visualisation. We can do this by applying the pixel-wise mean on all images of a domain. Furthermore, information about the variance within one domain can be extracted by applying standard deviation on a pixel basis.

3.2.2. Qualitative

In our qualitative analysis, we focus on domain shift based on texture. The pipeline consists of feature extraction and feature selection. The goal is to get two significant, uncorrelated features.

Feature extraction. Feature types can be classified into colour, texture, statistical and geometry features. We focus on texture, specifically grey level co-occurrence Matrix (GLCM), and statistical features.

A grey-level co-occurrence matrix (GLCM) is a histogram of co-occurring greyscale values at a given offset over an image and serves as a compact summary of the matrix. Features that are commonly extracted from GLCM are contrast, dissimilarity, homogeneity, ASM (ASM" value shows the strength of homogeneity, namely the pair correlation), energy, and correlation.

Statistical features are based on the whole image. This includes the mean, which can be assumed to be the mean brightness in a greyscale image.

Feature selection. Types of feature selection include removing features with low variance, univariate feature selection, recursive feature elimination, L1-based feature selection, tree-based feature selection and sequential feature selection.

In the related area of bioinformatics, where feature analysis and determining the significance of a feature is a common task, we can see the random forest classifier being a popular choice (Qi, 2012).

Redundant feature removal with correlation matrix. A correlation matrix helps in understanding which features are redundant. We remove correlating features in order to avoid overfitting, increase interpretability as feature importance will be incorrect with redundant features and reduce unnecessary dimensions.

The Pearson correlation measures the strength of the linear relationship between two variables. It has a value between -1 to 1, with a value of -1 meaning a negative linear correlation, 0 being no correlation, and + 1 meaning a positive correlation (Kotu and Deshpande, 2019).

Feature importance analysis with random rorest. In order to understand the importance of a feature, we can use a random forest classifier in combination with a metric, such as the Gini impurity, also known as mean decrease in impurity (MDI) or the permutation importance also known as mean decrease in accuracy (MDA).

A random forest classifier (Breiman, 2001) consists of an ensemble of decision trees. It incorporates feature selection and interactions naturally in the learning process. By using the dataset name as a class label, we can understand which features are most relevant for distinguishing between the datasets, hence, for which features the distribution shift is highest.

Gini impurity (Yuan et al., 2021) is useful in decision trees for calculating the purity of the branches of our tree. In an ideal case, each branch of the decision tree will be homogenous in that it contains a single class. If the branch is pure, this ideal case will be satisfied (Gini impurity = 0). Any Gini impurity score above 0 can be used to understand the homogeneity, or lack of, in the data on the branches.

The feature importance based on feature permutation can also be calculated, this being advantageous over impurity-based feature importance measures as they are not biased towards high-cardinality features. Cardinality refers to the number of distinct values that a feature can have, so high-cardinality features are those that have a large number of distinct values.

Feature significance analysis with univariate feature selection. In order to understand the significance of a feature, we can use univariate feature selection in combination with a metric, such as the p-value or the chi-squared score.

Univariate feature selection works by selecting the best features based on univariate statistical tests by removing all but the highest scoring features.

The p-value (Thiese et al., 2016) is a measure of the relationship between two groups of data. A low p-value means that there is likely a strong relationship between the groups of data, and that the null hypothesis - the statement that there is no association between groups - should be rejected. Inversely, a high p-value suggests there is likely no relationship between the groups of data, and that the null hypothesis should be accepted.

The chi-squared statistic (Plackett, 1983) is a common statistical test used to determine the significance of any association between two variables, doing so by comparing the observed frequencies of the variables with the frequencies which should be expected if they were independent. The chi-squared score is computed between each non-negative feature and class.

3.3. Transforms

All transforms used and compared in this work are stand-alone. They may be trained in their own pipeline, but do not get influenced by the final semi-supervised loss.

3.3.1. Fourier domain adaptation

Frequency decomposition has shown to promote content-style disentanglement where low frequencies are an approximation for style and high frequency are an estimate for the content.

With the help of the fast Fourier transform (FFT), the amplitude can be extracted as an estimate of image style (domain-specific) while the phase represents the image content (domain-invariant) (Yang and Soatto, 2020).

3.3.2. Multi-modal unsupervised image-to-image translation network

The multi-modal unsupervised image-to-image translation network (MUNIT) (Huang et al., 2018) is a neural network-based approach to disentangle content and style of two datasets in order to perform image-to-image translation. Given an image in the source domain, the goal is to learn the conditional distribution of corresponding images in the target domain, without having access to any paired images. The MUNIT architecture consists of two auto-encoders, one for each domain. Each auto-encoder has a latent space for content and one for style. Image-to-image translation is performed by swapping encoder-decoder pairs. A simplified diagram of a single content-style disentangled auto-encoder can be seen in Figure 6. For training, an adversarial loss (L_{GAN}) ensures, that the translated images are not able to be distinguished between the target and source domains. Furthermore, bidirectional reconstruction losses for image reconstruction (L_{rec}), content latent reconstruction (L_{c-rec}) and style latent reconstruction (L_{s-rec}) make sure that encoders and decoders are inverses.

The final loss function to be minimised is

$$L = L_{GAN} + \lambda_1 L_{rec} + \lambda_2 L_{c-rec} + \lambda_3 L_{s-rec} \quad (1)$$

3.4. Data Sampler

3.4.1. Mixed Batch Sampling

Due to the number of different datasets, a mixed batch sampling (MBS) approach is used. Similar to the two-stream batch sampling, the MBS approach samples n images from each data source for each epoch. This also helps with class imbalance as it naturally applies undersampling of all classes greater than the minority class (Anand et al., 2010).

3.5. Mean Teacher

The Mean Teacher approach, introduced by Tarvainen and Valpola (2017), makes use of two equal

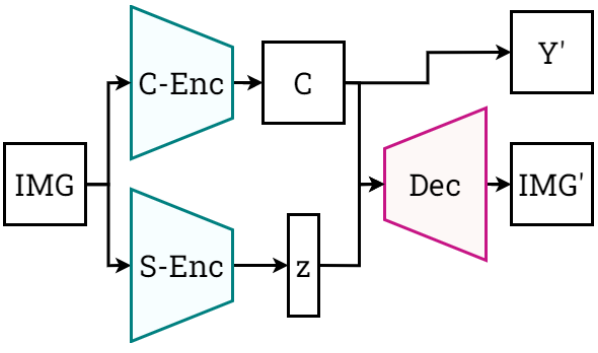


Figure 6: Auto-encoder of the MUNIT with a content and a style encoder, feeding two disentangled latent spaces. z denotes the style encoding, while C denotes the content encoding. IMG' is the reconstructed image. Y' is an additional output that may be used for segmentation purposes.

models, a student model and a teacher model, which are trained simultaneously. The student model is trained with both labelled and unlabelled data, while the teacher model is updated by the exponential moving average of the student model's parameters.

A very commonly used model for instance segmentation in the medical imaging domain is the U-Net (Ronneberger et al., 2015). It consists of an encoder-decoder structure with skip connections. The encoder captures hierarchical features, while the decoder recovers spatial information. Skip connections help to preserve fine details and contextual information.

In order to construct a U-Net, a backbone is needed, for instance, the EfficientNet b0 (Tan and Le, 2019). It is designed to balance model depth, width, and resolution for optimal performance by following a compound scaling method that uniformly scales these dimensions to create a highly efficient and effective model.

3.6. Updating the student network

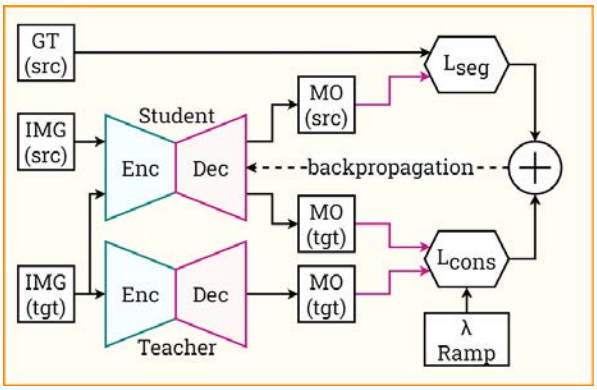


Figure 7: Updating the student network through backpropagation which is part of the Mean Teacher. IMG are the labelled source and unlabelled target images. MO are the model outputs. GT is the ground truth of the source domain. Only the student model gets backpropagated through.

The student network is updated using the standard backpropagation algorithm on the labelled data. See Figure 7.

3.6.1. Loss overview

For training the student network, we combine a supervised loss term (L_{seg}) focusing on the labelled data and a consistency loss term ($L_{cons} * \lambda_{ramp}$) focusing on the unlabelled data.

$$L = L_{seg} + L_{cons} * \lambda_{ramp} \tag{2}$$

L_{seg} is the dice loss between the student's predictions of all labelled images and their ground truth.

L_{cons} is the MSE/L2 loss between the student's predictions and the cleaned teacher's predictions. Here, only unlabelled images are taken into account.

λ_{ramp} is a weighting factor, that weights the consistency higher the more the epochs progress.

An additional loss term $L_{certain}$ is introduced in the last experiment.

3.6.2. Segmentation loss function

Dice loss, also known as the Sørensen-Dice coefficient or F1-score loss, is a commonly used loss function for segmentation tasks, including instance segmentation. It measures the similarity or overlap between the predicted segmentation masks and the ground truth masks.

Dice loss is particularly useful in scenarios where the foreground objects of interest are small in proportion to the background. It helps address the class imbalance issue that often arises in segmentation tasks, where the background pixels heavily outnumber the object pixels.

Dice Loss:

$$L_{seg}(mo_s, gt) = 1 - \frac{2 \times |mo_s \cap gt|}{|mo_s| + |gt|} \tag{3}$$

$|mo \cap gt|$ denotes the number of pixels in the intersection between the predicted "model output" mask and the ground truth mask, $|mo| + |gt|$ represents the total number of pixels in the predicted and ground truth masks, respectively.

3.6.3. Consistency loss function

The consistency loss can be defined as the discrepancy between the model's predictions on the original and perturbed inputs. It can be formulated using various metrics, such as mean squared error, Kullback-Leibler divergence, or cosine similarity.

We calculat it as the mean squared error between the predictions of the teacher and the student models on the unlabeled data, and it serves as a regulariser to encourage the student model to produce similar predictions as the teacher model. This helps to improve the generalisation performance of the model by reducing overfitting to the labelled data.

$$\text{MSE} = \frac{1}{n} \times \sum (mo - gt)^2 \tag{4}$$

The final consistency loss function is defined as:

$$L_{cons}(mo_s, mo_t) = \frac{1}{n} \times \sum (mo_s - mo_t)^2 \tag{5}$$

For the last experiment, we use the Dice score instead of the mean squared error in order to calculate the loss between the student's model output and the cleaned pseudo label generated by the teacher model.

3.6.4. Accounting for uncertainty

Epistemic (systematic) uncertainty describes what the model does not know because training data was not appropriate. Epistemic uncertainty is due to limited data and knowledge.

Aleatoric (statistical) uncertainty is the uncertainty arising from the natural stochasticity of observations.

We can quantify the amount of uncertainty in an entire probability distribution using the Shannon entropy. The entropy is used as an additional loss for the pipeline accounting for uncertainty, which is the third approach.

$$L_{certain}(mo_{tta}) = -\sum mo_{tta} * \log(mo_{tta}) \quad (6)$$

In order to calculate entropy, we can use test-time augmentation. An image is augmented with different noise sources, and put through a model. The entropy is calculated between the model outputs. See Figure 23. Furthermore, we generate a combined mask for the consistency term, where we use the mean of the model predictions, then take the biggest area and fit an ellipse on it. In this way, we want to reduce artefacts as well as push the network to make more circular predictions.

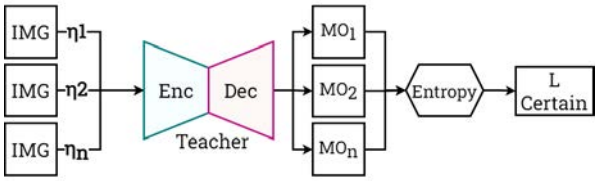


Figure 8: Certainty calculation, where img is one single image, n is the noise source, and MO is the model output.

3.6.5. Epoch-based weighting

According to the temporal ensembling approach for semi-supervised learning, proposed by (Laine and Aila, 2016), a ramp-up function for weighting the supervised loss and the consistency loss component is needed. They state, that we cannot rely on the consistency loss from epoch zero, hence a way of performing a warm start is needed. One way of achieving an increasingly stronger weighted consistency loss term is introducing a ramp-up function. At the beginning, the supervised loss dominates, until the point of the ramp-up length (r_{max}) is reached, then both losses contribute equally. Commonly, exponential ramp-up (see Figure 9), or a simple linear function are applied.

The exponential ramp-up is defined as

$$\lambda_{ramp}(t, r_{max}) = \exp(-\frac{1}{2} \times (1 - \frac{t}{r_{max}})^2) \quad (7)$$

where t is the current epoch and r_{max} is the ramp-up length.

3.6.6. Back-propagation

Only the student model backpropagates the error, for which an optimiser and additionally a learning rate scheduler are applied. For instance, Loshchilov and Hutter (2016) introduced stochastic gradient descent with warm restarts (SGDR) also known as cosine annealing learning rate scheduler with warm restarts, where warm restarts are simulated by scheduling the learning rate. They state, that SGDR may also make learning rate selection easier since the annealing and restarts consider a range of learning rate values.

3.7. Updating the teacher network

3.7.1. Exponential Moving Average

Instead of averaging predictions as it is done in temporal ensembling, the mean teacher approach averages the models' weights. Precisely, the teacher model's weights are updated using the Exponential Moving Average (EMA) of the student model's weights. The teacher model is updated at each iteration of the training process (Tarvainen and Valpola, 2017).

The EMA is calculated as follows:

$$EMA(w, t) = (\mathrm{w} \times \alpha) + (EMA(w, t-1) \times (1 - \alpha)) \quad (8)$$

where w is the weight of the student model at epoch t. EMA(w, t-1) is the EMA of the weight at the previous iteration. α is the decay factor.

3.8. Post-processing

Due to the anatomy of the human eye, which includes a single optic disc, we filter the image using the largest region. This approach enables us to effectively remove artefacts and enhance the quality of the image.

3.9. Evaluation Matrices

Different scores are available for segmentation and the medical field (Hicks et al., 2022).

Qualitative metrics include recall, precision and F1-score.

Recall also referred to as the sensitivity or True Positive Rate (TPR), indicates the proportion of correctly classified positive samples. It is computed as the ratio of correctly classified negative samples to all samples classified as negative.

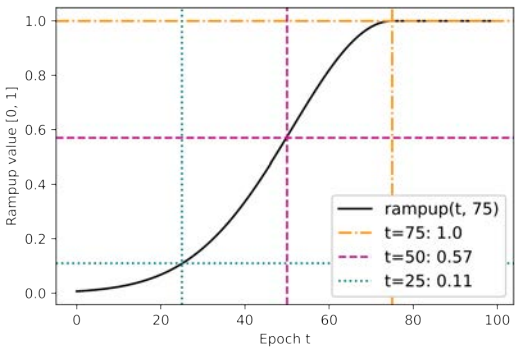


Figure 9: Exponential ramp-up function, with an example ramp-up length of 75.

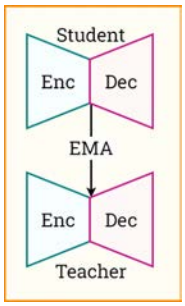


Figure 10: Updating the teacher network with exponential moving average (EMA) which is part of the Mean Teacher.

Precision is calculated as the ratio of correctly classified samples to all samples from a specific class.

F1-score is a measure that combines precision and recall into a single metric. It ranges from 0 to 1, where 1 represents maximum precision and recall values and 0 represents zero precision and/or recall.

For quantitative analysis, we visualise masks.

4. Results

4.1. Domain shift analysis

4.1.1. Visually

In the mean image (Figure 11), we can see, that the optic disc in the OCT images are dark generally with vessels being darker on the top and the bottom of the optic disc, this can be explained due to anatomy. The translated fundus image shows artefacts caused by the Fourier domain adaptation. The black circular border around fundus images is the source of error.

For the standard deviation (Figure 12) the lighter area indicates higher variability. We can see that CIRRUS image has more variability within the image, the MUNIT was able to partially follow the pattern. This may be due to the random styles applied to the PLEX content.

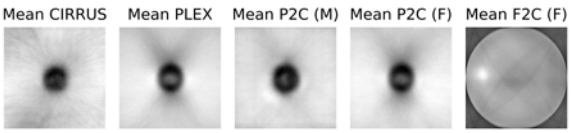


Figure 11: Mean within the domain. Image-to-image translation is performed by MUNIT (M) or Fourier domain adaptation (F).

4.1.2. Qualitative

For the quantitative analysis, we use the CIRRUS and PLEX data for feature selection. Using the correlation matrix, we can see the following: The contrast and dissimilarity are positively correlated, the homogeneity is negatively correlated. Furthermore, there is a high correlation between ASM and energy.

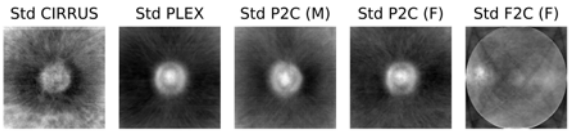


Figure 12: Standard deviation within the domain. Image-to-image translation is performed by MUNIT (M) or Fourier domain adaptation (F).

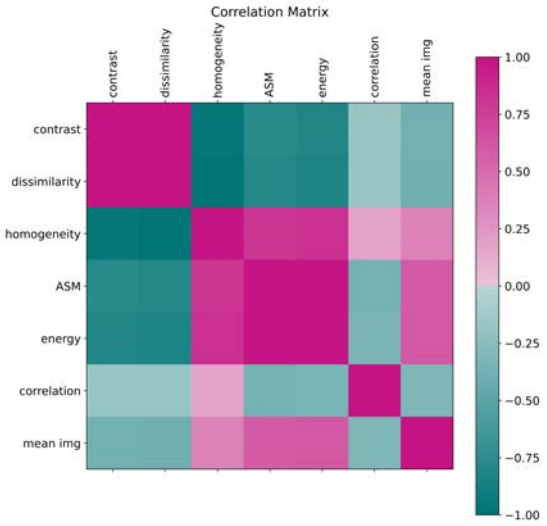


Figure 13: Correlation of features for PLEX and CIRRUS.

We choose four features, contrast, correlation, energy and the mean image. Out of these, contrast is the strongest feature. This can be seen in both the random forest feature analysis as well as the univariate feature selection. See Figures 14 and 15.

The final features chosen in order to track the domain shift are contrast as well as mean image.

We can see, that the CIRRUS Glaucoma and CIRRUS Normal have no domain shift. Between the CIRRUS and PLEX data, especially the contrast is varying. PLEX images within the domain also show a great variety in contrast.

The transformed images can be interpreted as follows.

The MUNIT generally produces pictures with low contrast, lower than both PLEX and CIRRUS. However, the contrast ratio within the image domain is smaller.

FDA managed to account for the contrast shift.

Fundus images are generally darker, which can be interpreted by the mean image value. Both, the fundus and the MUNIT-translated PLEX images are rather low in contrast.

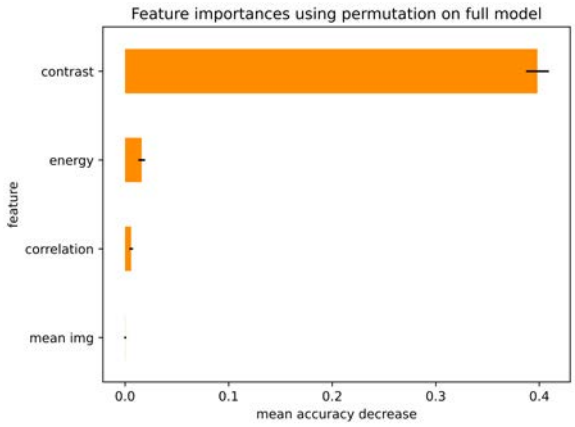


Figure 14: Feature importance using random forest.

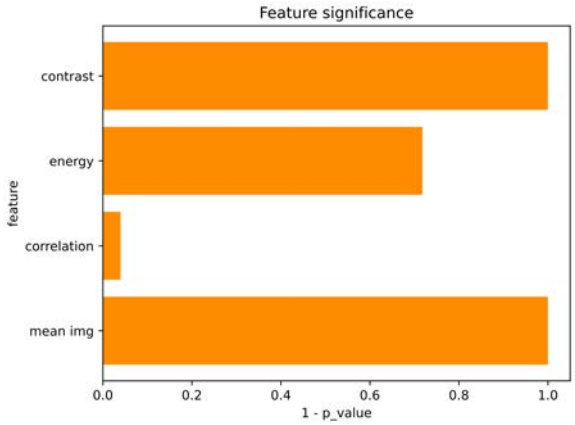


Figure 15: Feature importance using p-value. As the p-value shows significance with values towards zero, we use 1-p value. The bigger the bar, the higher the significance.

4.2. Transforms

4.2.1. MUNIT: Fundus-to-CIRRUS

We chose the MUNIT due to the fact that we do not need image pairs of the source and target domain. Augmentations include cropping, flipping and normalisation. We chose an image size of 256x256, and a style vector size of 8. The MUNIT was trained with 3 channel images. The Fundus-to-CIRRUS MUNIT was trained for 400 iterations with CIRRUS images as one domain and fundus images as the other domain. Visually it can be seen, that the generated images have artefacts. Furthermore, the results are not useful for training, since the optic disc of the newly generated images is not in the centre. The results can be seen in Figures 17 and 18. The progression of training the MUNIT can be seen in the appendix.

4.2.2. MUNIT: PLEX-to-CIRRUS

The PLEX-to-CIRRUS MUNIT was trained for 200 iterations with CIRRUS images as one domain and

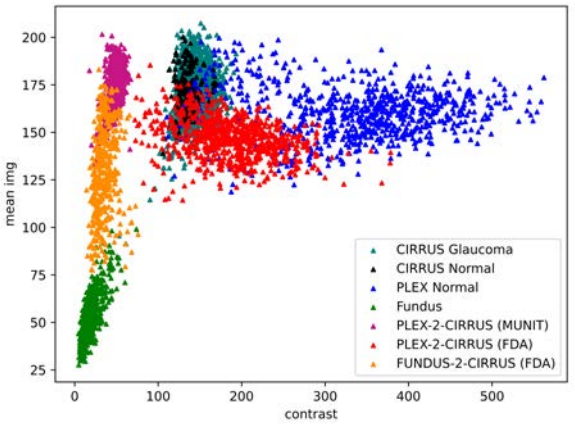


Figure 16: Overview of different datasets and their domain shift.

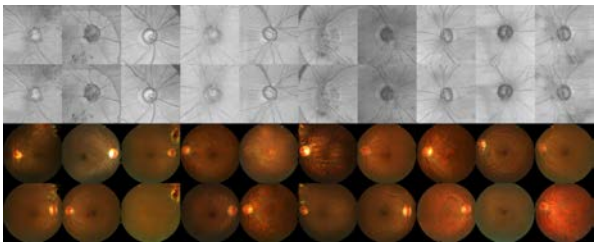


Figure 17: CIRRUS-to-fundus with the MUNIT. The first row shows the CIRRUS samples from the testset. The second row shows the reconstructed image. The two last rows show two sets of newly generated fundus images.

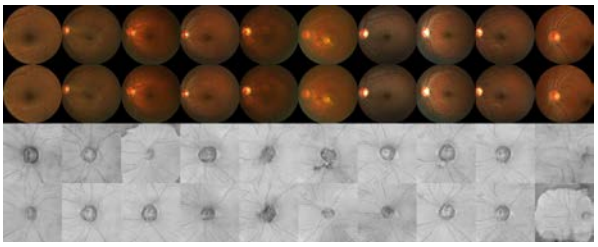


Figure 18: Fundus-to-CIRRUS with the MUNIT. The first row shows the fundus samples from the testset. The second row shows the reconstructed image. The two last rows show two sets of newly generated CIRRUS images.

PLEX images as the other domain. The hyperparameters are equal to the previous experiment. The results can be seen in Figures 19 and 20.

4.2.3. FDA: Fundus/PLEX-to-CIRRUS

A hyperparameter for the Fourier domain adaptation approach was chosen as beta=0.5. Visual results can be seen in Figures 21 and 22.

4.3. Mean Teacher setup and hyperparameters

We want to give an overview of design choices and hyperparameters across our experiments.

Augmentations include resizing and cropping the image to 128x128, random vertical and horizontal flips,

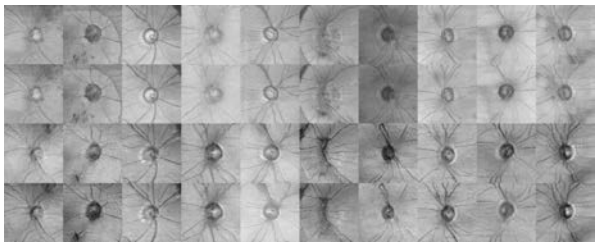


Figure 19: CIRRUS-to-PLEX with the MUNIT. The first row shows the CIRRUS samples from the testset. The second row shows the reconstructed image. The two last rows show two sets of newly generated fundus images.

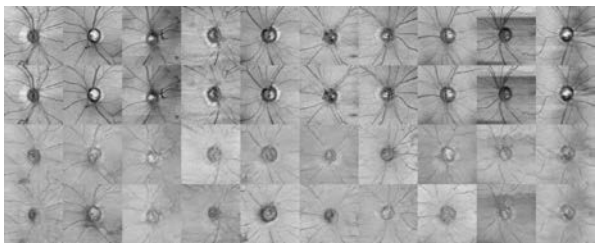


Figure 20: PLEX-to-CIRRUS with the MUNIT. The first row shows the fundus samples from the testset. The second row shows the reconstructed image. The two last rows show two sets of newly generated CIRRUS images.

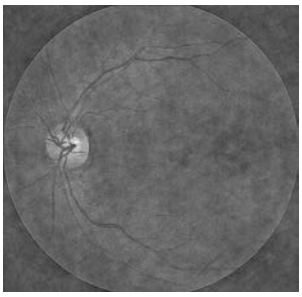


Figure 21: An example for fundus-to-CIRRUS translation using Fourier domain adaptation.

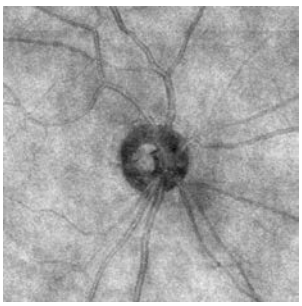


Figure 22: An example for a PLEX-to-CIRRUS translation using Fourier domain adaptation.

a random combination of brightness, contrast, gamma and sharpness factors, random blurring, inversion, and finally min-max normalisation. These transforms were chosen due to the domain shift.

We use the Efficientnet b0 backbone, pre-trained on imagenet for proof of concept as it is computationally less expensive.

We use one input channel (greyscale) and two output neurons.

We use a ramp-up length (r_{max}) of 300. Experiments with a shorter ramp-up length resulted in unstable learning.

The network is training for 300 epochs, using early-stopping based on the f-score. If there is no improvement for 10 epochs, training is stopped.

In the context of this model, one epoch is defined as a complete pass of the downsampled dataset through the model. This means, that only n images are used from each dataset, n being the size of the smallest dataset. The batch size for each dataset is 4. Hence, the final batch size is 12.

The learning rate is determined by the Cosine Annearling Learning Rate Scheduler with warm restarts. The base learning rate is 0.001, with a minimum learning rate of 0.00001. The momentum is 0.9.

The ema decay value remains as the default value, 0.999.

4.4. Uncertainty

In order to account for uncertainty, which may be linked to artefact segmentation, we change the pipeline as follows. First, we add different noise to an image four times in order to get 4 augmented versions of an image. We put it through the network. Then we take the mean of the outputs. In the last instance we use the biggest area and fit an ellipse around it. The MSE loss in replaced by a Dice loss, and the combined and cleaned prediction is used as a pseudo mask.

4.5. Segmentation Pipeline

We chose to proceed with the Fourier domain adaptation as an image-to-image translation technique because of two reasons. Firstly, FDA works well on both PLEX and fundus data. Furthermore, the alignment according to the contrast feature gave better results for FDA than MUNIT.

The final model is chosen with the help of early stopping, when there is no increase of f-score for 10 epochs, the training is stopped. The f-score is calculated based

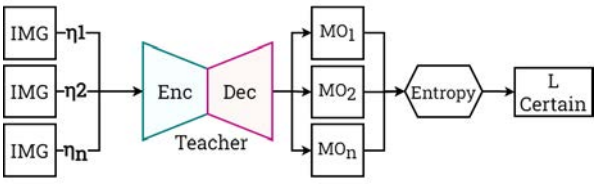


Figure 23: Uncertainty visualised. Student is the output of the student model. Teacher is the output of the teacher model. N1 is the probability map of a single noise image put through the teacher network. Clean is the cleaned mean image of the four noise images. Img is the reference image.

on the validation set (PLEX). Models applying Fourier domain adaptation, transform images in training, as well as a validation set. The final F1 score, precision and recall are calculated based on the CIRRUS testset, which contains 50 manually annotated masks.

The results of the different experiments can be seen in Table 1.

Name	Epoch (vF1)	F1	Prec	Rec
MT	114 (0.897)	0.879	0.935	0.844
F+MT	107 (0.884)	0.895	0.916	0.887
F+MT+C	87 (0.881)	0.901	0.916	0.897

Table 1: Results of the Mean Teacher (MT) baseline and its variations. Validation F-score (vF1), F-score (F1), Precision (Prec), Recall (Rec), Fourier domain adaptation (F), certainty (C).

5. Discussion

5.1. Domain shift analysis

With the help of techniques for exploratory data analysis, we were able to understand the datashift between the PLEX and the CIRRUS data.

Due to the high accuracy of the random forest classifier, we can see, that chosen features differ greatly between two domains.

While the CIRRUS data has more variance within its images, especially the illumination is less consistent. Furthermore, CIRRUS images are brighter than PLEX images.

5.2. Transforms

5.2.1. Interpretation of MUNIT results between CIRRUS and fundus data

In order to understand the unwanted results of the MUNIT in this experiment, we need to look into the inference stage of the network. Generally, only the two generators are used to generate the new samples. The generator is used as an encoder and as a decoder. In the encoding direction, we translate an image into its content and style. In the decoding direction, we translate an image from a content and a style representation back into an image. The implementation allows us to either use the style extracted from a target image, or use a random vector.

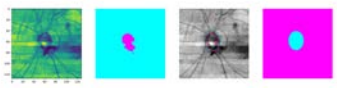


Figure 24: Example result of the baseline.

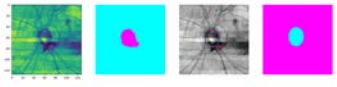


Figure 25: Example result of FDA.

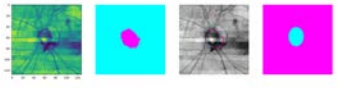


Figure 26: Example result of FDA with a cleaned mask.

See Figure 27 for a visualisation between PLEX (in this example Fundus) and CIRRUS.

The image-to-image translation with the MUNIT worked in a way, that a CIRRUS image got translated to a fundus image and vice versa. However, the network was not able to correlate the optic disc of the fundus and the CIRRUS data. This may be due to multiple reasons:

The MUNIT network is not designed for this task. It does image-to-image translation, hence generates a typical CIRRUS image from a fundus image. It does not do image registration-related tasks, which would have been somewhat the goal.

The target and source domain would need to have somewhat aligned content. This means, for example the optic disc needs to be centered for both domains.

Both points are caused cause the domain shift is too high, the network may not find the correlation of the optic discs in both domains. The problem is not the colour (style), but the shifted anatomy (content). Due to the dark macula which is centered in the fundus image, the MUNIT may connect the content of optic disc in the CIRRUS with the macula in the fundus image.

Furthermore, the limited amount of data may play a role. Additionally, the problem may be solved by a longer training duration.

In conclusion, the easiest way of solving the mismatch is the alignment of the optic discs.

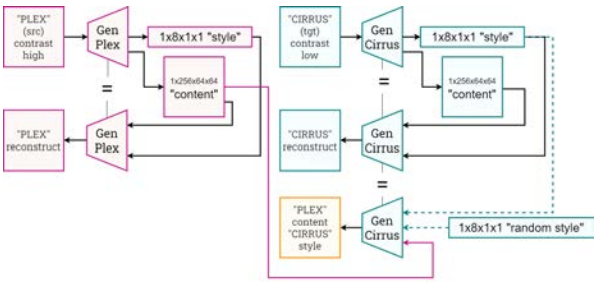


Figure 27: MUNIT at inference. Example based on PLEX (or fundus) and CIRRUS. There are only two generators used for inference, the PLEX generator and the CIRRUS generator. Both are used for encoding and decoding throughout the inference. The style vector may either be extracted from a (random) source image, or randomly generated.

5.2.2. Interpretation of MUNIT results between CIRRUS and PLEX data

Due to normalisation during the training, brightness problems appear in the result data. The images are

slightly blurry which can be solved by training for a longer period of time as well as increase the image size.

5.2.3. Interpretation of FDA results

Fourier domain adaptation gave both, in CIRRUS-to-PLEX as well as CIRRUS-to-fundus, decent results. The limitation is, that only the low-frequency space is exchanged and hence no natural artefacts are generated.

5.3. Segmentation Pipeline

The performance of all models are similar, this may be due to the simplicity of the task.

Generally, dark areas are mistaken to be part of the optic disc, hence artefacts which are common in CIRRUS images appear to be predicted as the foreground class.

6. Conclusions

We introduce a pipeline for optic disc segmentation incorporating image-to-image translation as a standalone data augmentation technique and semi-supervised learning to learn from labelled/transformed and unlabelled data.

Multiple experiments were carried out.

Transforms were applied to account for the domain shift. This includes, next to basic transforms such as inversion, blurring and change in intensity, also image-to-image techniques such as MUNIT and Fourier domain adaptation.

As for semi-supervised learning, we used a mean-teacher network, using both a supervised loss for the labelled data and a consistency constraint for the unlabelled data. Dependent on the experiment, we also perform uncertainty estimation as well as pseudo mask cleaning of the teacher's model output.

6.1. Future Work

A range of steps can be taken to potentially improve the segmentation results. The hyperparameters for the Fourier domain adaptation filter can be learned. A shape-based loss function based on the Hausdorff distance may reduce artefacts. Additionally, more investigation on the $L_{certainty}$ term should be made. Adversarial learning seems like a promising, or even superior, alternative, as widely used in literature. More tests with the MUNIT should be made, potentially using a different image size, longer training, as well as aligning optic discs of fundus and CIRRUS images.

Acknowledgments

I want to say thank you to my supervisor Damon for being patient with my chaotic thoughts as well as for being motivating, my supervisor Prof. Jacky for helping with eye-related questions, my lab head Prof. Leo for assigning a super interesting project to me, my mum Gabi for trying to understand what I am doing, my friend James for checking my English and also trying to understand my work, my friend Nisma for always being supportive and doing online work sessions with me even though we have a 6-hour time difference, my friend Chen Chen for explaining her work to me, my former colleague Michi for basically teaching me PyTorch from scratch, my side-project colleague Moritz for introducing me to uncertainty and out-of-distribution detection, and last but not least, my current lab colleagues Valentina, Yvonne, Susan and Padmini for being my debug rubber ducks.

References

Anand, A., Pugalenthi, G., Fogel, G.B., Suganthan, P., 2010. An approach for classification of highly imbalanced data using weighting and undersampling. Amino acids 39, 1385–1391.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A., 2019. Mixmatch: A holistic approach to semi-supervised learning. Advances in neural information processing systems 32.

Blum, A., Mitchell, T., 1998. Combining labeled and unlabeled data with co-training, in: Proceedings of the eleventh annual conference on Computational learning theory, pp. 92–100.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Chen, C., Ouyang, C., Tarroni, G., Schlemper, J., Qiu, H., Bai, W., Rueckert, D., 2020. Unsupervised multi-modal style transfer for cardiac mr segmentation, in: Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges: 10th International Workshop, STACOM 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Revised Selected Papers 10, Springer. pp. 209–219.

Chen, C., Qin, C., Ouyang, C., Li, Z., Wang, S., Qiu, H., Chen, L., Tarroni, G., Bai, W., Rueckert, D., 2022. Enhancing mr image segmentation with realistic adversarial data augmentation. Medical Image Analysis 82, 102597.

Choi, J., Kim, T., Kim, C., 2019. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6830–6840.

Chua, J., Baskaran, M., Ong, P.G., Zheng, Y., Wong, T.Y., Aung, T., Cheng, C.Y., 2015. Prevalence, risk factors, and visual features of undiagnosed glaucoma: the singapore epidemiology of eye diseases study. JAMA ophthalmology 133, 938–946.

Fu, H., Li, F., Orlando, J., Bogunovic, H., Sun, X., Liao, J., Xu, Y., Zhang, S., Zhang, X., 2020. Adam: Automatic detection challenge on age-related macular degeneration. IEEE Dataport .

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. The journal of machine learning research 17, 2096–2030.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020. Generative adversarial networks. Communications of the ACM 63, 139–144.

Grandvalet, Y., Bengio, Y., 2004. Semi-supervised learning by entropy minimization. Advances in neural information processing systems 17.

Guo, L., Moss, S.E., Alexander, R.A., Ali, R.R., Fitzke, F.W., Cordeiro, M.F., 2005. Retinal ganglion cell apoptosis in glaucoma is related to intraocular pressure and iop-induced effects on extracellular matrix. Investigative ophthalmology & visual science 46, 175–182.

Heijl, A., Bengtsson, B., Oskarsdottir, S.E., . Prevalence and severity of undetected manifest glaucoma: results from the early manifest glaucoma trial screening. Ophthalmology 120.

Hicks, S.A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M.A., Halvorsen, P., Parasa, S., 2022. On evaluation metrics for medical applications of artificial intelligence. Scientific Reports 12, 5979.

Huang, X., Liu, M.Y., Belongie, S., Kautz, J., 2018. Multimodal unsupervised image-to-image translation, in: Proceedings of the European conference on computer vision (ECCV), pp. 172–189.

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134.

Kazemi, H., Iranmanesh, S.M., Nasrabadi, N., 2019. Style and content disentanglement in generative adversarial networks, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. pp. 848–856.

Kotu, V., Deshpande, B., 2019. Chapter 4 - classification, in: Kotu, V., Deshpande, B. (Eds.), Data Science (Second Edition). second edition ed.. Morgan Kaufmann, pp. 65–163.

Laine, S., Aila, T., 2016. Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242 .

Liu, M.Y., Breuel, T., Kautz, J., 2017. Unsupervised image-to-image translation networks. Advances in neural information processing systems 30.

Liu, M.Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., Kautz, J., 2019. Few-shot unsupervised image-to-image translation, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 10551–10560.

Liu, X., Sanchez, P., Thermos, S., O'Neil, A.Q., Tsaftaris, S.A., 2022. Learning disentangled representations in the imaging domain. Medical Image Analysis , 102516.

Loshchilov, I., Hutter, F., 2016. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 .

Miyato, T., Maeda, S.i., Koyama, M., Ishii, S., 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE transactions on pattern analysis and machine intelligence 41, 1979–1993.

Ouyang, C., Kamnitsas, K., Biffi, C., Duan, J., Rueckert, D., 2019. Data efficient unsupervised domain adaptation for cross-modality image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22, Springer. pp. 669–677.

Pascolini, D., Mariotti, S.P., 2012. Global estimates of visual impairment: 2010. British Journal of Ophthalmology 96, 614–618.

Plackett, R.L., 1983. Karl pearson and the chi-squared test. International statistical review/revue internationale de statistique , 59–72.

Qi, Y., 2012. Random forest for bioinformatics, in: Ensemble machine learning: Methods and applications. Springer, pp. 307–323.

Qin, C., Li, W., Zheng, B., Zeng, J., Liang, S., Zhang, X., Zhang, W., 2023. Dual adversarial models with cross-coordination consistency constraint for domain adaption in brain tumor segmentation. Frontiers in Neuroscience 17.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer. pp. 234–241.

Saito, K., Saenko, K., Liu, M.Y., 2020. Coco-funit: Few-shot unsupervised image translation with a content conditioned style encoder, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, Springer. pp. 382–398.

Sánchez-Migallón, M.C., Valiente-Soriano, F.J., Nadal-Nicolás, F.M., Vidal-Sanz, M., Agudo-Barriuso, M., 2016. Apoptotic retinal ganglion cell death after optic nerve transection or crush in mice: delayed rgc loss with bdnf or a caspase 3 inhibitor. Investigative ophthalmology & visual science 57, 81–93.

Senjam, S.S., 2020. Glaucoma blindness–a rapidly emerging noncommunicable ocular disease in india: Addressing the issue with advocacy. Journal of Family Medicine and Primary Care 9, 2200.

Sharma, P., Sample, P.A., Zangwill, L.M., Schuman, J.S., 2008. Diagnostic tools for glaucoma detection and management. Survey of ophthalmology 53, S17–S32.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L., 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems 33, 596–608.

Spaide, R.F., Fujimoto, J.G., Waheed, N.K., Sadda, S.R., Staurenghi, G., 2018. Optical coherence tomography angiography. Progress in retinal and eye research 64, 1–55.

Sun, B., Feng, J., Saenko, K., 2017. Correlation alignment for unsupervised domain adaptation. Domain adaptation in computer vision applications , 153–171.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR. pp. 6105–6114.

Tan, N.Y., Friedman, D.S., Stalmans, I., Ahmed, I.I.K., Sng, C.C., 2020. Glaucoma screening: where are we and where do we need to go? Current opinion in ophthalmology 31, 91–100.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems 30.

Thiese, M.S., Ronna, B., Ott, U., 2016. P value interpretations and considerations. Journal of thoracic disease 8, E928.

Toldo, M., Maracani, A., Michieli, U., Zanuttigh, P., 2020. Unsupervised domain adaptation in semantic segmentation: a review. Technologies 8, 35.

Volpi, R., Namkoong, H., Sener, O., Duchi, J.C., Murino, V., Savarese, S., 2018. Generalizing to unseen domains via adversarial data augmentation. Advances in neural information processing systems 31.

Wang, H., Shen, T., Zhang, W., Duan, L.Y., Mei, T., 2020. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV, Springer. pp. 642–659.

Weih, L.M., Nanjan, M., McCarty, C.A., Taylor, H.R., 2001. Prevalence and predictors of open-angle glaucoma: results from the visual impairment project. Ophthalmology 108, 1966–1972.

Weinreb, R.N., Aung, T., Medeiros, F.A., 2014. The pathophysiology and treatment of glaucoma: a review. Jama 311, 1901–1911.

Yang, Y., Soatto, S., 2020. Fda: Fourier domain adaptation for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4085–4095.

Yow, A.P., Tan, B., Chua, J., Husain, R., Schmetterer, L., Wong, D., 2021. Segregation of neuronal-vascular components in a retinal nerve fiber layer for thickness measurement using oct and oct angiography. Biomedical Optics Express 12, 3228–3240.

Yuan, Y., Wu, L., Zhang, X., 2021. Gini-impurity index analysis. IEEE Transactions on Information Forensics and Security 16, 3154–3169.

Zeng, G., Lerch, T.D., Schmaranzer, F., Zheng, G., Burger, J., Gerber, K., Tannast, M., Siebenrock, K., Gerber, N., 2021. Semantic consistent unsupervised domain adaptation for cross-modality medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24, Springer. pp. 201–210.

Zhang, Y., Deng, B., Jia, K., Zhang, L., 2020. Label propagation with augmented anchors: A simple semi-supervised learning baseline for unsupervised domain adaptation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV, Springer. pp. 781–797.

Zhao, S., Li, B., Yue, X., Gu, Y., Xu, P., Hu, R., Chai, H., Keutzer, K., 2019. Multi-source domain adaptation for semantic segmentation. Advances in Neural Information Processing Systems 32.

Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE international conference on computer vision, pp. 2223–2232.

Appendix

Progress of training a MUNIT

In order to visualise the progress of the MUNIT, we choose the example of CIRRUS-to-fundus image translation as well as CIRRUS-to-PLEX, since this translation is more intuitive for understanding the MUNIT results. It becomes very clear, that the content alignment is not working in the CIRRUS-to-fundus translation, for instance, multiple optic discs are present in the fundus results. We use the train data for visualisation. A more detailed interpretation is present in each image caption.

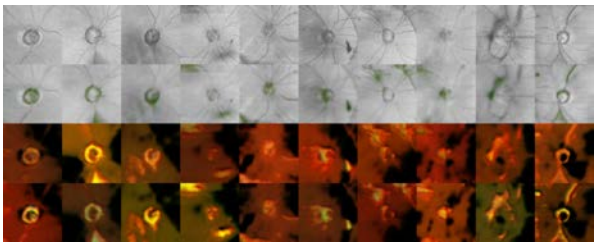


Figure 28: Iteration 1: The reconstructed image in row two shows green artefacts. The fundus image represents what we would have wanted to receive as a final result. Unfortunately, the quality is rather bad and the vessels are bright.

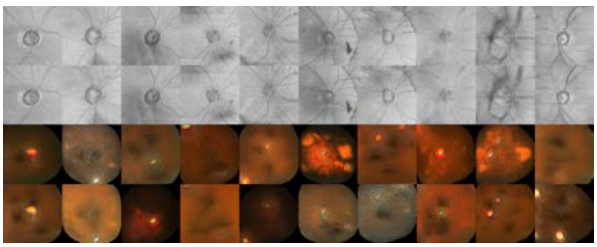


Figure 29: Iteration 100: After 100 iterations, the reconstructed image is marginally blurry. The result images tend to have the optic disc in the middle. Due to the fact that the CIRRUS image is way more zoomed in, the optic disc in the result fundus image appears too small.

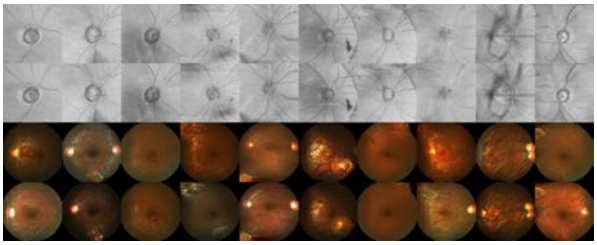


Figure 30: Iteration 200: The MUNIT performs well on the reconstruction task. Also, the general anatomy of a fundus image is achieved. Occasionally, two optic discs are present in one image, either mirrored horizontally or vertically. No optic disc is in the centre anymore. Artefacts, especially on the borders of the fundus image are present.

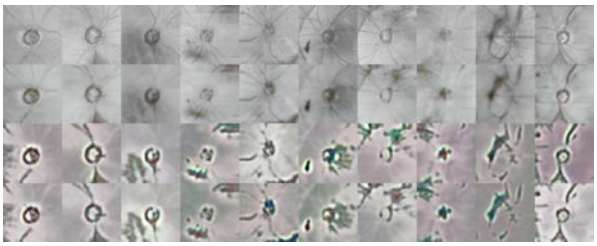


Figure 31: Iteration 1: The reconstructed image, similar to the CIRRUS-to-fundus translation, shows artefacts. Generally, the same effects as in the other experiment are present.

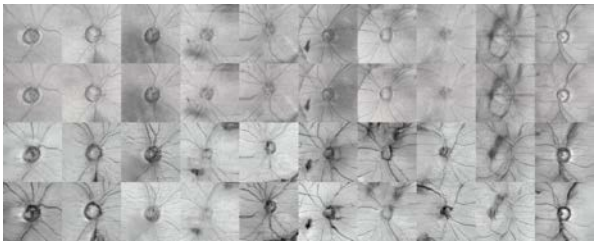


Figure 32: Iteration 100: Different to the CIRRUS-to-fundus experiment, we can already see in iteration 100, that the MUNIT understood the different style. It correctly takes the content of the CIRRUS image and combines it with the style of the PLEX image. Note that this is not the direction we need in our work (PLEX-to-CIRRUS). Translating CIRRUS to another domain makes it just visually better understandable, how the MUNIT is working.

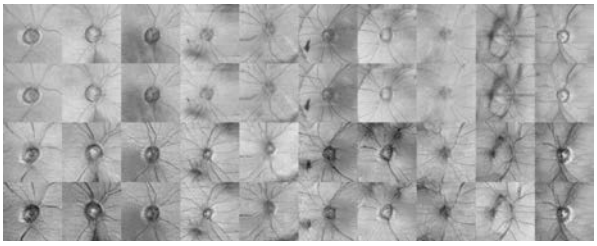
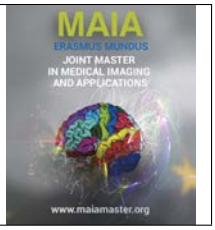


Figure 33: Iteration 200: Similarly to iteration 100, the PLEX shows a higher contrast. Also, the natural artefacts in OCT images are present in both target and source domain which is a good sign, as we need natural artefacts to increase the robustness of the semi-supervised approach.

Medical Imaging and Applications

Master Thesis, June 2023



A full pipeline to analyse lung histopathology images

Lluis Borràs Ferrís, Niccoló Marini, Henning Müller

MedGIFT, University of Applied Sciences Western Switzerland, Sierre (HES-SO), TechnoArk 3, 3960 Sierre

Abstract

Histopathology images are the gold standard to diagnose most different cancer types. These images are usually analysed by a pathologist through optical inspection of the glass slides in modern microscopes. With a limited number of pathologists and an increasing number of biopsies and resections performed, this work aims to explore modern alternatives to alleviate the workload of pathologists. Digital Pathology concerns the acquisition and management of glass slides and producing whole slide images (WSI) to be inspected on the computer. The workflows in hospitals are shifting to a new domain where pathologists examine the WSIs on the computer instead of traditional microscope observation. Computational Pathology emerges thanks to the slides' digitalisation and therefore, the creation of big WSIs datasets and it aims to develop computer-aided diagnosis (CAD) systems helping to reduce pathologists' workload in tasks such as cancer segmentation and classification. Lung cancer holds the top position as the primary cause of cancer-related deaths and has the second-highest incidence rate worldwide. Precise classification among the different lung cancer subtypes is a crucial task that determines target treatments that improve the survival of patients with such malignancy. A CAD system is proposed based on self-supervised pre-training and multiple instance learning (MIL) training to classify lung WSIs into four classes, three major cancer subtypes, small cell lung cancer (SCLC), non-small cell lung cancer, adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC) and normal tissue (NL). Trained in two different private datasets with 2,226 WSIs the model obtain an AUC of 0.8558 ± 0.0051 and a weighted f1-score of 0.6537 ± 0.0237 in the 4-class classification on the test set. Moreover, to evaluate the generalisation capability of the model, it was tested on the public TCGA dataset with LUAD and LUSC subtypes and obtained an AUC of 0.9433 ± 0.0198 and a weighted f1-score of 0.7726 ± 0.0438.

Keywords: Self-supervised, Weakly-supervised, MIL, CAD, Lung Cancer, Histopathology, LUAD, LUSC, SCLC, Automatic labels, Machine Learning, Deep Learning

1. Introduction

Morphological Histopathology is widely regarded as the reference standard for the diagnosis of most cancers (Tornillo and Franco, 2022). Currently, in most hospitals, pathologists visually examine slides using microscopes without the aid of modern technologies. The typical workflow in a histopathology laboratory involves collecting either a biopsy or a tissue resection (part of an organ) for examination. For resections, macroscopic inspection is conducted to select the tissue segments that will undergo microscopic examination by the pathologists. This differs from biopsy scenarios where all extracted tissue is prepared for microscopic examination. The tissue undergoes various stages including dehydration with formalin and other chemical agents, and subsequent embedding in paraffin to prepare it for cutting. A microtome cuts the tissue into slides, which are typically 3 to 5 µm thick. Various stains are applied to differentiate the different structures present on the slide. The most used stain is Hematoxylin & Eosin (H&E). Hematoxylin produces a blue-purple colour and stains nucleic acids, whereas Eosin produces a pink colour and stains basic structures. In a typical tissue sample, the nuclei are stained darker due to the presence of DNA, while the cytoplasm and extracellular matrix display varying degrees of pink staining. Histopathology images include several tissue structures, ranging from microscopic entities (such as single-cell nuclei) to macroscopic components (such as tumour solid mass). (Fis-

cher et al., 2008)

The number of biopsies and resections collected worldwide has been increasing over the years due to several factors such as increasing screening strategies to diagnose cancer before its symptoms are present and to deliver a final diagnosis that would determine the best-personalized therapy plan. In contrast, the number of pathologists is not increasing equally with a consequent workload on the sector. A study performed by Märkl et al. (2021) explored the ratio of pathologists per number of inhabitants in Europe, the USA and Canada. In Europe, on average, there is a ratio of one pathologist per 32,018 inhabitants varying from 14,309 on Island to 63,028 in Poland. Switzerland, the USA and Canada have one pathologist per 35,355, 20,658 and 25,325 inhabitants, respectively. Therefore, the scientific community aims to explore modern alternatives to alleviate the workload of pathologists.

1.1. Digital pathology and computational pathology

A Whole Slide Image (WSI) is a digitized slide that is scanned at high-resolution and stored in a multi-scale (pyramidal) format as shown in Figure 1. Digital pathology is becoming increasingly integrated into some hospitals, with an additional step in the workflow involving the digitisation of glass slides through the use of automated digital pathology scanners that offer magnification equivalent to a microscope. Observation of slide images in such cases usually involves a hybrid workflow, whereby, depending on the urgency of the matter, inspections are conducted either through traditional microscopy or visualizing the scanned WSI directly on specialized screens for improved resolution. The acquisition of a WSI typically occurs at x40 magnification level, resulting in images to over 100,000 pixels in each dimension at the highest resolution level, with a pixel size of 0.25 μm. Consequently, more public and private datasets are available with histopathological images (Marini et al., 2021a).

Computational pathology aims to develop automatic algorithms to analyse WSIs unleashing the power of digital pathology. Most of these algorithms are currently based on Machine learning (ML), specifically on Deep Learning (DL) algorithms to improve the accuracy and efficiency of cancer diagnosis. By using large amounts of data from digital pathology images, DL algorithms can learn to identify and classify different types of cancer and provide additional insights to aid pathologists in making a final diagnosis. ML-powered image analysis can also help automate repetitive and time-consuming tasks, such as tumour segmentation and cell counting, enabling pathologists to focus on more complex cases and improving overall diagnostic accuracy (Abels et al., 2019).

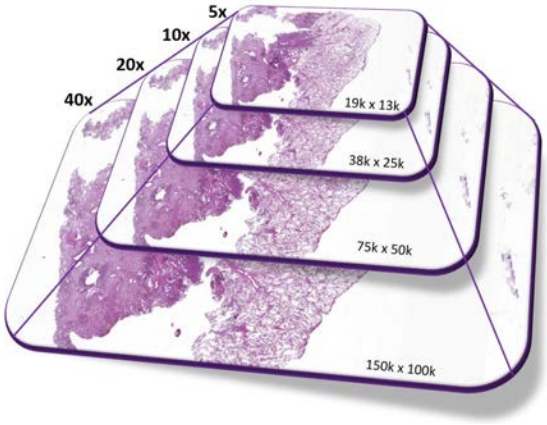


Figure 1: Example of digitized whole slide image (WSI) scanned at 40x (0.25 μm/pixel) as high-resolution and stored in a multi-scale (pyramidal) at four different magnification levels.

1.2. Lung cancer

Lung cancer currently exhibits the highest mortality rate among all cancer types with an 18.0 age-standardized mortality rate (ASR) per 100,000 inhabitants including all ages in 2020 (Ferlay et al., 2020). Additionally, it has the second-highest incidence rate, with an estimated 2,206,771 new cases reported in the same year. In Europe, lung cancer holds the highest mortality rate, with a 22.6 ASR per 100,000 inhabitants including all ages, and the third highest incidence rate, with an estimated 477,534 new cases in 2020 (Ferlay et al., 2021; Sung et al., 2021). The lung cancer locations include the whole lung, the bronchus, their parts, the mediastinum, the thoracic lymph nodes, the pleura, and the pulmonary lymph nodes (Travis et al., 2011).

The initial step in diagnosing lung cancer involves performing chest radiography on patients who exhibit symptoms associated with either local or systemic effects of the tumour. In cases where radiography indicates positive results, a biopsy is performed on the area with abnormal lung findings (American Cancer Society). In accordance with ICD-11 guidelines, pathologists carefully examine the biopsy samples and issue a report that includes the lung cancer subtype (if a tumour is present) and TNM staging (Lababede and Meziane, 2018). Histopathology remains the gold standard for cancer diagnosis and is critical in determining a patient's prognosis and in identifying appropriate surgical and/or treatment interventions.

Lung cancer is classified into two primary groups: non-small-cell lung cancer (NSCLC) and small-cell lung cancer (SCLC). NSCLC is further categorized into three subtypes: adenocarcinoma (LUAD), squamous cell carcinoma (LUSC), and large-cell lung carcinoma. Among these, LUAD is the most common subtype and accounts for 50% of all NSCLC cases. Adenocarcinoma also has six subtypes, including acinar, papillary, mucosal invasive, lepidic, micropapillary, and solid. Of

all lung cancer cases, 80% are classified as NSCLC, while the remaining 20% are classified as SCLC (Goldstraw et al., 2011; Van Meerbeeck et al., 2011). Accurate identification of the distinct categories and subcategories is essential due to the varied prognosis for the patient, and treatment options can vary significantly depending on the cancer subtype, which is a critical step in the diagnostic process that can profoundly impact patient survival.

SCLC is the lung cancer type with the worst prognosis as it has a high capacity for rapid metastasis, resulting in a low survival rate of 31% in the localized stage and 2% in the disseminated stage after 5 years. In contrast, for non-small cell lung cancer (NSCLC), the size of the primary tumour is a crucial factor affecting the survival rate in stage I, while the number of malignant nodules (N1) is the main factor determining the survival rate in stage II. The survival rate in stage IV is approximately 50%, while it is only 1% in stage IV (Kumar et al., 2018).

Computational pathology could play a very important role in pathologists' workflow. By designing Computer-aided diagnosis (CAD) systems using ML algorithms to classify different lung cancer subtypes. The development of these CAD tools could potentially help pathologists in the analysis of WSIs and reduce their workload with an increasing number of biopsies performed on the hospitals (Otálora et al., 2021).

1.3. Contributions

In this work, an algorithm is proposed to perform a 4-class classification task among the 3 most prevalent cancer types SCLC, LUAD and LUSC and healthy tissue as shown in Figure 2. The model is pre-trained in a self-supervised algorithm and trained using a weakly-supervised learning strategy by training a Multiple Instance Learning (MIL) model using both, labels automatically extracted from the pathologist's reports and manual labels at WSI-level annotated by a pathologist.

This work presents the following contributions.

- An innovative pipeline is proposed that combines self-supervised pre-training and weakly-supervised training using MIL for the classification task of lung cancer between four different classes, the most prevalent three cancer types (SCLC, LUAD and LUSC) and normal tissue.
- A comparison of the model trained on a private cohort with manual annotation from an expert pathologist and automatic labels obtained from the reports. Additionally, to evaluate the performance of using self-supervised learning to pre-train the model, its performance is compared with the model pre-train on ImageNet in the downstream classification task.

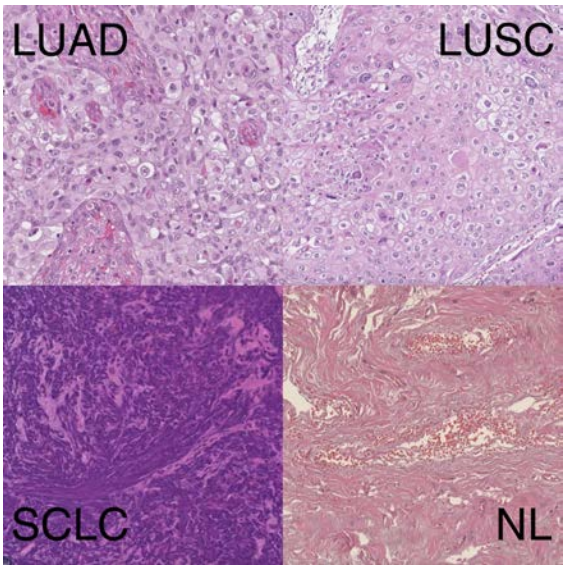


Figure 2: Representatives patches of the four different subtypes used to train the model. LUAD: Non-small-cell adenocarcinoma, LUSC: Non-small-cell squamous cell carcinoma, SCLC: Small-cell lung cancer, NL: Normal tissue.

- The models are tested on The Cancer Genome Atlas (TCGA) public dataset to analyse their generalisation capabilities.
- An interpretability metric of the performance of the self-supervised model is presented by analysing the features extracted from 384 patches composed of cells, glands or stroma, by using a Uniform Manifold Approximation and Projection (UMAP) for dimension reduction and plotting the results to evaluate if the model is capable to separate the three different patch-types in different clusters.

2. State of the art

The diagnosis of lung cancer is a key factor in the survival time. Currently, the conventional method of diagnosing lung cancer in most hospitals is through the examination of histopathology slides by pathologists using a microscope. This method is considered the gold standard for cancer diagnosis, but it can be a complex and time-consuming process, as the morphological differences among lung cancer subtypes are subtle. The correct classification of lung cancer subtype is critical for determining the most appropriate surgical and treatment options for the patient (Coudray et al., 2018).

Recent advances in computational pathology and DL techniques have demonstrated the potential in improving tumour histopathology evaluations. WSIs are images digitized at high-resolution, with dimensions ranging from 10,000 to over 150,000 pixels. Due to their large size, WSIs are commonly divided into patches,

Table 1: Review of state-of-the-art methods in the classification of histopathological lung cancer whole slide images highlighting the different training strategies for learning and the number of whole slide images (WSI) used on the specific datasets. LUAD: Non-small-cell lung adenocarcinoma, LUSC: Non-small-cell lung squamous cell carcinoma, NL: Normal, SCLC: small-cell lung carcinoma, PTB: pulmonary tuberculosis, OP: Organizing pneumonia, AUC: Area under de ROC curve, TCGA: The Cancer Genome Atlas, ICGC: International Cancer Genome Consortium KMC: Kyushu Medical Centre, MH: Mita Hospital, TCIA: The Cancer Imaging Archive, DPGFLCD: Department of Pathology of the Georges François Leclerc Cancer Center in Dijon, UHC: University Hospital of Caen, TMUH: Taipei Medical University Hospital, WFH: Taipei Municipal Wanfang Hospital, SHH: Taipei Medical University Shuang-Ho Hospital, SYSU: First Affiliated Hospital of Sun Yat-sen University, SZPH: Shenzhen People's Hospital dataset.

Paper	Training stratagy	Dataset	Subtypes	Results	Preprocessing
Coudray et al. (2018)	Fully-supervised	TCGA	567 LUAD, 609 LUSC or 459 NL	AUCs of 0.993 tumour vs. NL, 0.950 LUAD vs. LUSC, 0.968 3-class	Patch
Yu et al. (2020)	Fully-supervised	TCGA ICGC	427 LUAD 457 LUSC 87 LUAD 38 LUSC	AUC LUAD vs. LUSC 0.927 ± 0.004, AUC LUAD vs. LUSC 0.842 ± 0.011	Patch
Wang et al. (2020)	Semi-supervised (coarse annotations)	WSI private dataset TCGA	390 LUAD 361 LUSC 120 SCLC, 68 NL	Accuracy of 0.973 TCGA: AUC 0.820, accuracy 0.820	Patch
Kanavati et al. (2020)	Weakly-supervised	4,054 KMC, 500 MH, 680 TCGA 500 TCIA	Lung carcinoma and non-neoplastic carcinoma	AUCs 0.975 KMC, 0.974 MH 0.988 TCGA, and 0.981 TCIA	Patch
Le Page et al. (2021)	Fully-supervised	DPGFLCD, UHC TCGA	66 nonLUSC 66 LUSC, 45 nonLusc 20 LUSC, 30 nonLUSC 30 LUSC	Accuracy 0.85 DPGFLCD UHC: 0.81 AUC, TCGA: AUC 0.78	Patch
Chen et al. (2021)	Weakly-supervised	TMUH, WFH, SHH TCGA	3,876 LUAD 1,088 LUSC, 2,039 NL TCGA: 532 LUAD and 512 LUSC	AUCs 0.9594 LUAD 0.9414 LUSC TCGA: 0.8950 LUAD 0.8990 LUSC	Resize
Lu et al. (2021)	Weakly-supervised	TCGA	55 LUAD and 55 LUSC	AUC of 0.902 ± 0.016 for lung	Patch
Yang et al. (2021)	Fully-supervised	741 SYSU1, 318 SYSU2 212 SZPH and 422 TCGA	LUAD, LUSC, SCLC, PTB, OP and NL	AUCs 0.970 SYSU1, 0.918 SYSU2, 0.963 SZPH and 0.978 TCGA	Patch
Kanavati et al. (2021)	Fully-supervised	1,723 KMC, 500 MH, and 905 TCGA	LUAD, LUSC SCLC and NL	AUCs 0.94 - 0.99 in LUAD, LUSC, SCLC and neoplastic vs. non-neoplastic	Patch
Chen et al. (2022)	Self-supervised Weakly-supervised	TCGA	10,678 33 cancer types 1,008 LUAD and LUSC	Self-supervised pre-trained on 10,678 WSI AUC of 0.952 ± 0.021 LUAD vs LUSC	Scaling

as current graphics processing units (GPU) cannot handle the entire WSI at its original size. Traditional deep learning approaches require local labels for each patch, which are time-consuming and expensive to create in the medical field (Marini et al., 2021b).

The lack of large, annotated, datasets and data heterogeneity are still open challenges in computational pathology. In a typical fully-supervised training pixel-wise annotations are needed to train the model which is a time-consuming task for pathologists and, therefore, very expensive. Several methods are proposed to solve these problems such as semi-supervised and weak-supervised learning as shown in Table 1.

2.1. Fully-supervised learning

Fully-supervised learning (Strong supervision) relays on manual pixel-level annotations to train the deep learning models. That means that for every patch on the WSI, a label must be provided to train the model. Usually, it requires a pathologist or group of pathologists to provide these manual pixel-level annotations, which are expensive and a very time-consuming task. These fully-supervised models achieve the best performances and are the most widely strategy used in the state-of-the-art models in lung cancer subtype classification. Using patch-level annotations, strong labels, Yang et al. (2021) achieves a micro-average area under the curve (AUC) of 0.970 among 3 different cancer types (LUAD, LUSC and SCC), two cancer mimics, pulmonary tuberculosis (PTB) and Organizing pneumonia (OP) and normal tissue examples using a private dataset (SYSU1) and an AUC of 0.978 in TCGA public dataset. On the other hand, Coudray et al. (2018) with the same training strategy achieves an AUC of 0.968 in a 3-class classification between LUAD, LUSC and Normal. Classifying between LUSC and non-LUSC in a smaller dataset Le Page et al. (2021) obtains an AUC of 0.81 in a private external dataset and 0.78 in the TCGA. Yu et al. (2020) obtains an AUC of 0.927 in the training TCGA dataset and an AUC of 0.842 in an independent dataset. Finally, Kanavati et al. (2021) obtains AUCs between 0.94 to 0.99 in the test set in 3 different datasets in four various binary classification problems between LUAD, LUSC, SCLC and neoplastic tissue vs. non-neoplastic tissue.

2.2. Semi-supervised learning

In the case of semi-supervised algorithms, instead of providing pixel-level annotations for all the data, the strategy is to have a big portion of the dataset with weak labels (labels at only WSI-level) using an automatic algorithm to annotate the data and a smaller percentage of pixel-level annotations to train the model. This strategy is more convenient but still needs some effort from a pathologist to manually annotate some WSIs on the dataset. Using coarse annotations from pathologists Wang et al. (2020) achieves an accuracy of 0.973 in a private dataset and an AUC of 0.820 in the TCGA public dataset.

2.3. Weakly-supervised learning

The line of research that investigates how to best use image-level diagnostic labels, is known as weakly-supervised learning. Without pixel-level annotations, weak supervision models approach the training of a model only using WSI-level annotations which better

replicates the real scenario if a pathologist provides only one diagnosis per image. These WSI-level annotations are noisy by nature because only a small portion of the patches are representative of the label. To solve this problem among all the weakly-supervised algorithms, MIL is the state-of-the-art. To represent the bags (WSIs) in the MIL frameworks, two different strategies to aggregate the instance-level features into a bag-level representation are studied. These strategies aim to capture the key characteristics of the lung tissue samples and differentiate between the different cancer types and normal cells.

- **Instance-level aggregation:** One approach, is to build an instance-level classifier that returns scores for each patch. Then the individual scores are aggregated by MIL pooling (such as max pooling or average pooling). This pooling operation summarizes the information within each patch, capturing essential features associated with different cancer types.

- **Embedding-level aggregation:** Alternatively, the instances are mapped to a low-dimensional embedding. Afterwards, MIL pooling is used to obtain a bag representation independent of the number of patches in each bag.

Lu et al. (2021) proposed an algorithm called CLAM which is a deep-learning-based weakly-supervised method that obtains an AUC of 0.956 ± 0.02 in the TCGA public dataset to discriminate between LUAD or LUSC. Kanavati et al. (2020) suggested also a weakly-supervised training using only WSI-level diagnoses on a dataset of 9,662 lung cancer WSIs. This method achieves an AUC of 0.959 and 0.941 for LUAD and LUSC on the testing set, respectively.

2.4. Transfer learning

Transfer learning approaches leverage pre-trained models that have been trained on extensive datasets like ImageNet or Instagram 1-Billion. The primary objective is to take advance of models that have already acquired a feature representation of a sizable image dataset. Consequently, the classifier layers of the network can be retrained, or in some cases, specific layers of the model can be unfrozen to learn representative features from the images in a new dataset. This process involves training the model on the targeted dataset to perform the new classification task. (Cheplygina et al., 2019). All of the works, with the exception of the last one that uses self-supervised learning, presented in Table 1 take advantage of this strategy to load a pre-trained network trained on ImageNet and train only the classifier and in the specific lung cancer classification task (Deng et al., 2009).

2.5. Self-supervised learning

In recent years, unsupervised representation algorithms have gained prominence in the field of computer vision. Instead of relying on pre-training models with weights from ImageNet, these approaches aim to pre-train the models using the dataset's own images, constructing tokenized dictionaries for unsupervised learning. For example, in natural language processing (NLP), tokenisation is the process of breaking down the text into smaller inputs, such as words, called tokens. A tokenized dictionary would contain these individual tokens as its entries, allowing for efficient lookup and analysis of specific words within the dictionary. In the computer vision domain building these dictionaries is an open challenge since the data exists in a high-dimensional space.

He et al. (2020) addressed this challenge by introducing Momentum Contrast (MoCo), a technique that constructs dynamic, large, and consistent dictionaries using contrastive loss. In their work, they demonstrate that MoCo effectively narrows the gap between unsupervised and supervised representations in computer vision tasks such as object detection and segmentation, employing widely recognized datasets like PASCAL VOC and COCO.

Additionally, Chen et al. (2020a) proposed a straightforward algorithm for contrastive learning. Through their work on SimCLR, they highlighted the significance of data augmentation composition, the incorporation of a learnable nonlinear transformation between the representation and the contrastive loss and larger batch sizes (4k - 8k batch size) together with more training steps. With these three important findings, they enhanced model effectiveness and achieved a new state-of-the-art on the ImageNet dataset.

Building upon these findings from SimCLR, the researchers at Facebook AI Research introduced MoCo v2 (Chen et al., 2020b), which incorporated more aggressive data augmentation and an MLP projection head exhibited improved performance compared to the work of the Google research team on the ImageNet dataset. What's more, they show that with MoCo v2 is possible to process a large set of negative samples without requiring large training batches and consequently powerful GPUs. In contrast, Moco v2 can run on a typical 8-GPU machine.

Furthermore, Dehaene et al. (2020) demonstrated in their work that leveraging MoCo v2 in a self-supervised learning framework effectively closed the gap between weakly-supervised and fully-supervised learning using histopathology images from the Camelyon16 dataset.

Chen et al. (2022) uses a different strategy to pre-train the self-supervised model using a new approach with the potential of transformers called DINO and applying it to histopathology data (Caron et al., 2021). They used DINO for pre-training with 10,678 WSIs from 33

different cancer types collected from the public TCGA dataset. Afterwards, they train a MIL model, using weak labels, for a binary classification task on 1,008 WSIs of LUAD and LUSC achieving an AUC of 0.952 ± 0.021.

2.6. Pre-processing strategy

In the context of the aforementioned discussion, WSIs are images characterized by an immense number of pixels, rendering it unfeasible to process them in their original size due to limitations posed by GPU hardware. In the field of computational histopathology, researchers have devised two distinct strategies to overcome this challenge. The prevalent approach, employed by a majority of researchers working with histopathology images (as observed in 5 out of the 6 studies detailed in Table 1), involves partitioning the images into smaller patches. These patches are subsequently utilized to train an ML model, enabling the model to learn a downstream task.

Alternatively, Chen et al. (2021) pursued an alternative methodology, which involved resizing the WSIs to dimensions of 21,000x21,000. In their study, they directly trained a model on the resized WSIs to perform classification tasks using weak labels distinguishing between LUAD and LUSC. By implementing this approach, they achieved an AUC of 0.9594 for LUAD and 0.9414 for LUSC on a private dataset, while obtaining corresponding AUC values of 0.8950 and 0.8990 for LUAD and LUSC, respectively, on the TCGA public dataset.

In the case of Chen et al. (2022) instead of preprocessing the WSIs patching or resizing them, they take advantage of the potential of transformers to scale through different stages to learn representable features of these high-resolution images from lower to higher patch-level resolutions to capture information from individual cells to tissue microenvironment.

3. Material and methods

This section is dedicated to describe the complete methodology employed in this work. Describing the datasets employed for training and testing, the complete pipeline, the experimental set-up choice for both, training the self-supervised and the weakly-supervised models, the evaluation criteria and a final section describing a visualisation tool for the classification results in more detail than only a simple label per WSI.

3.1. Datasets

The self-supervised model is trained using all the data from Azienda Ospedaliera per l'Emergenza Cannizzaro Catania (AOEC) with a total of 1,354 WSIs. The goal is to extract high-representative features specific to histopathological lung data. All the WSIs were preprocessed and divided into patches at 10x magnification resulting in a total of 2,950,251 images.

Table 2: Overview of the dataset composition. The datasets include lung images from digital pathology laboratories in Azienda Ospedaliera per l'Emergenza Cannizzaro Catania (AOEC) and Radboud University Medical Centre (RUMC) used for training and testing. The training dataset is divided into train and validation using 5-fold cross-validation. Additionally, the model is tested on The Cancer Genome Atlas (TCGA) public dataset. SCLC: small-cell lung carcinoma, LUAD: Non-small-cell lung adenocarcinoma, LUSC: Non-small-cell lung squamous cell carcinoma, SKET: Semantic Knowledge Extractor Tool.

Source	SCLC	LUAD	LUSC	Normal	Total labels	Total images
Training dataset: automatic weak labels (SKET):						
AOEC	51	715	367	164	1,297	1,225
Training dataset: manual weak labels (Pathologist):						
AOEC	53	601	353	237	1,244	1,225
Training dataset from two different private datasets:						
AOEC	53	601	353	237	1,244	1,225
RUMC	0	297	205	499	1,001	1,001
Total	53	898	558	736	2,245	2,226
Testing private datastets:						
AOEC	17	16	9	14	46	46
RUMC	0	29	18	45	92	92
Total	17	45	27	59	138	138
Testing public dataset:						
TCGA	0	530	506	0	1,036	1,036

Two different MIL models are trained following two different approaches as shown in Table 2.

- A first model is trained on data provided by AOEC. The goal is to compare the results obtained in the same dataset with labels coming from two different sources. Automatic labels using the Semantic Knowledge Extractor Tool (SKET) (Marchesin et al., 2022; Marini et al., 2022) extracting labels directly from the pathologist's reports and manual labels provided by an expert pathologist.

- A second model is trained using WSIs from two different private datasets AOEC and Radboud University Medical Centre (RUMC). The objective is to study the improvement of the model with more heterogeneity data coming from two different hospitals and to observe if training with more data will also help to develop a more accurate model.

The first model is tested on the private datasets in both scenarios, using only a set of WSIs only from the AOEC dataset with a set of 46 WSIs and on a separate set of 138 WSIs from both hospitals. The purpose is to compare the performance in data from the same dataset and data coming from Catania and Radbound hospitals. The model trained with WSIs from the two hospitals is tested only on this second set of 138 unseen WSIs. Additionally, both models were tested on The Cancer Genome Atlas (TCGA) public dataset composed of 1,036 LUAD and LUSC WSIs (Albertina et al., 2016; Kirk et al., 2016) from 5 different centres in the USA (Washington University, University of Pittsburgh/UPMC, University

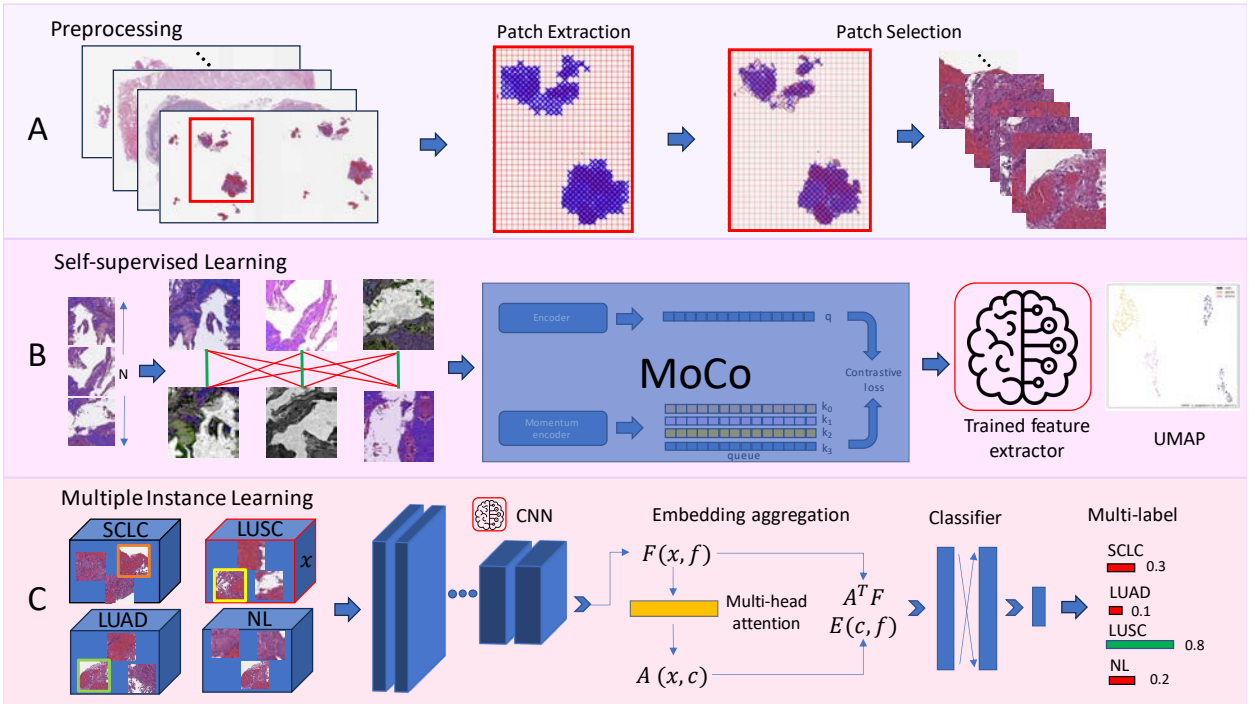


Figure 3: Complete Pipeline to train the lung cancer subtype classification model. A: Preprocessing of the whole slide images (WSI) to extract the patches used for training. B: Self-supervised learning to pre-train a feature extractor that captures high-level concepts directly from the histopathological lung patches using Momentum Contrastive Learning (MoCo). C: For each WSI, the features extracted, using the self-supervised model described in step B, are loaded to train a Multiple Instance Learning (MIL) model for the lung cancer classification task among, Non-small-cell adenocarcinoma (LUAD), Non-small-cell squamous cell carcinoma (LUSC), Small-cell lung cancer (SCLC) and, Normal tissue (NL), using a multi-label strategy. UMAP: Uniform Manifold Approximation and Projection, CNN: Convolutional Neural Network, x: number of patches per WSI, f: feature vector, c: number of classes.

of North Carolina, Lahey Hospital & Medical Center and Roswell Park).

The three different datasets are composed only of WSIs stained with H&E. The training datasets from AOEC and RUMC are imbalanced due to the characteristics of digital pathology workflows. With a higher number of WSIs with Normal tissue (negative biopsies and resections areas without malignancy) and LUAD, being the most prevalent subtype among the lung cancer ones, followed by LUSC and with fewer examples, SCLC, being the less prevalent one.

3.2. Pipeline

Figure 3 provides a comprehensive overview of the pipeline developed for classifying histopathology images of the lung into four distinct classes. First, all the WSIs are preprocessed to extract the patches used for training. Once all WSIs in the dataset are patched, these images are utilized for pre-training a self-supervised model based on MoCo v2 (Chen et al., 2020b). The objective is to train a feature extractor specific to the histopathology lung data. using the feature vectors from the previous step, the MIL model employs weak labels to train the classification model, to classify between the three most prevalent lung cancer types (LUAD, LUSC, SCLC) as well as normal WSIs.

3.2.1. Preprocessing

The pre-processing stage is composed of two components. The initial task involves extracting patches from each WSI. Each WSI is divided into multiple patches, ranging from a few hundred to thousands, depending on the WSI's size and the amount of tissue contained in each slide. The subsequent task involves selecting patches from the extracted set that contain representative tissue information from the respective WSI.

Patch Extraction: The initial pre-processing step is shown in Figure 4.A, it involves dividing each whole slide image (WSI) into patches to meet the GPU requirements for training the model at a specific magnification level. To accomplish this task, the PyHIST tool is chosen (Muñoz-Aguirre et al., 2020). This process involves three main steps:

1. Extraction of a mask that effectively separates the foreground (tissue) from the background content of the WSI.
2. Creation of a grid of non-overlapping tiles overlaid on the mask, followed by an evaluation to determine whether each tile belongs to the foreground or background.
3. Selection of patches by choosing tiles that fall within the extracted mask at the desired magnification level.

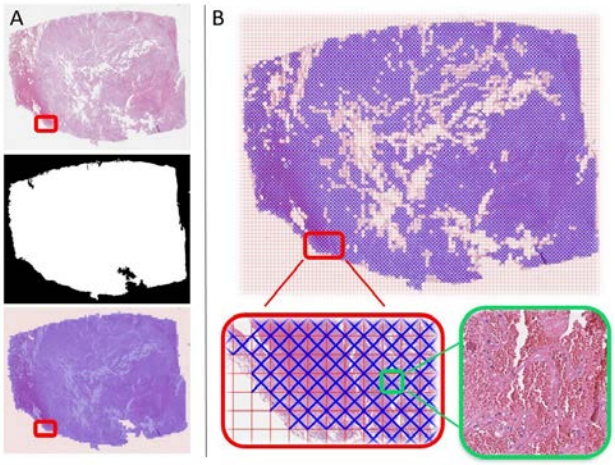


Figure 4: Pre-processing steps for one whole slide image (WSI). A: Patch extraction: by generating a mask and selecting the patches that fall into this mask. B: Patch selection: To remove wholes, patches with not enough information and errors from the extracted patches.

After consultation with an expert pathologist and considering the trade-off between hardware space and optimal patch resolution for the lung classification task, all the WSIs are downsampled and extracted patches at a 10x magnification level, with a tile size of 256x256. Based on the dataset characteristics, a downsampling factor of two is applied for images saved at a maximum magnification of 20x and a downsampling factor of four for images saved at a maximum magnification of 40x. This allowed us to obtain all patches at the desired 10x magnification.

To extract the mask, the algorithm initially identifies the tissue edges by applying the Canny edge detector algorithm (Canny, 1986). This edge image effectively separates the tissue information from the background, emphasizing the image borders. Subsequently, a graph-based segmentation algorithm (Boykov and Kolmogorov, 2004; Felzenszwalb and Huttenlocher, 2004) is employed to generate the final mask, which completely separates the background from the tissue content. This is achieved by applying a min-cut/max-flow algorithm to the edge image.

Finally, the mask image is divided into a non-overlapping grid and the WSI is downsampled to the requested magnification level. The patches are then selected by matching the corresponding tiles that intersect with the mask.

Patch Selection: The patch extraction algorithm achieves a low number of true negatives but produces a significant number of false positives, as shown in Figure 4.A, in the bottom image. Upon careful inspection of these false positives, several common characteristics are identified:

- The WSIs may contain macro holes, and all WSIs contain numerous micro holes that the graph cuts segmentation cannot accurately identify as background. These patches are typically whitish.
- Some WSIs contain text that does not contain tissue information and can be considered potential confounders. These texts are written in black letters.
- Some extracted patches usually located in the borders of the wholes do not contain enough tissue information.

To address this problem, the second step in pre-processing involves refining the extracted patches through a patch selection process. The primary objective is to reduce the number of false positives by removing unnecessary patches that would introduce noise in the model training, while still preserving patches with significant tissue information as illustrated in Figure 4.B.

To achieve the patch selection, an additional phase is proposed in the pre-processing pipeline that filters out unnecessary patches after the initial patch extraction performed by PyHIST. Considering the characteristics of the false positive patches, this step involves computing the histogram for each extracted patch. From the histogram, only the bins above and below a specific threshold are counted. If the number of bins within this threshold exceeds 50% of the total number of pixels in the patch, it is considered to contain important tissue information and is retained. Contrarily, if the number of pixels between the two thresholds falls below 50% of the total number of pixels, the patch is discarded.

There are important differences among the WSIs, with some being brighter and others darker, as shown in 2. Furthermore, as mentioned earlier, the false positive patches predominantly consist of white and black areas. Utilizing the mask computed by PyHIST, the average pixel intensity is computed of the gray scale image for the pixels contained within the mask. Three different pairs of thresholds are set based on this average intensity: if the average intensity is below 155, between 155 and 180, or above 180, lower thresholds of 35, 40, and 45 are set, and upper thresholds of 210, 215, and 220, respectively. This approach allows us to discard the majority of whitish background patches and black patches belonging to annotations in the images.

3.2.2. Self-supervised learning

In our investigation, it is observed that employing a feature representation of the images, rather than simply relying on weights from a model trained on ImageNet, can potentially produce better results. In the field of lung cancer classification, state-of-the-art methods commonly utilize a frozen feature extractor that has been pre-trained on ImageNet. However, this is

identified as a potential limitation. To address this issue, an alternative approach is suggested where a feature extractor is trained specifically for histopathology lung images. Self-supervised algorithms aim to learn a stronger data representation, exploiting data its-self, without the need for annotations. (Chen et al., 2020b; Dehaene et al., 2020).

To validate our proposal, the results obtained are compared from two different feature extractors in the downstream task of lung cancer classification using weakly-supervised learning with MIL. The results on both, training the MIL model using the features extracted from the self-supervised model and using the same exact model but extracting features using the pretrained model with weights from ImageNet are compared. As backbone for the self-supervised model, the ResNet34 is implemented (He et al., 2016), loaded from the PyTorch framework (Paszke et al., 2019).

By leveraging recent advancements in self-supervised learning, a feature extractor is trained specifically on histopathology images of the lung. MoCov2 self-supervised architecture (Chen et al., 2020b) is chosen as a reference from the FAIR research group. The primary objective is to train an encoder using contrastive learning. The encoder learns to associate images within the dataset by performing a dictionary look-up task summarized in Figure 3.B.

Contrastive learning is a machine learning technique used for unsupervised representation learning. It aims to learn useful features by contrasting positive pairs (similar samples) against negative pairs (dissimilar samples). In contrastive learning, a model is trained to map similar examples closer together in the feature space while pushing dissimilar examples apart (Hadsell et al., 2006). To achieve it, we have a scenario with an encoded query, denoted as q, and a collection of encoded samples, represented as k0, k1, k2, ..., which serve as the keys in a dictionary. It is assumed that within this dictionary, there exists a single key (referred to as k+) that matches the query q. Therefore, q and k will be a positive pair if they are data-augmented versions of the same image and negative otherwise.

This type of unsupervised learning is performed using contrastive loss. The value of this function is minimized when the query q is similar to its positive key k+ but dissimilar to all other keys in the collection, which are regarded as negative keys for q. The similarity between the query and keys is measured using dot product, and specifically, a variant of contrastive loss known as InfoNCE is considered (Wu et al., 2018) and presented in Equation 1:

$$\mathcal{L}_{\text{InfoNCE}} = -\log\left(\frac{\exp(q \cdot k^{+}/\tau)}{\sum_{i=0}^{N}\exp(q \cdot k_i/\tau)}\right) \qquad (1)$$

where:

q is the encoded query, a data-augmented version of the inputted patch. k^{+} denotes the positive key, which is the matching data-augmented version of the same patch coming from the momentum encoder that should have high similarity to the query. k_i represents the negative keys, which are unrelated patches that should have low similarity to the query. τ is a temperature parameter that controls the sharpness of the distribution. It is usually set to a small positive value.

The numerator of the fraction computes the exponential of the dot product between the query and the positive key, divided by the temperature τ. This term measures the similarity between the query and the positive key after applying a temperature scaling. The denominator sums the exponential of the dot products between the query and all keys (positive and negative). This term represents the normalisation factor, ensuring that the resulting values are in the range [0, 1]. Taking the logarithm of the fraction and negating it gives the final InfoNCE loss value. By minimizing this loss, the model learns to differentiate between related and unrelated patches, discovering meaningful representations.

The dictionary is built by using images as inputs that are highly dimensional and discrete. By building the dictionary as a queue of patches allows updating the queue from patches from the above mini batches decoupling the dictionary size from the mini-batch size. The queue size can be set as a hyperparameter. The queue represents a sampled subset of all data by enqueuing the last mini-batch and removing the oldest mini-batch in the queue. Then together with this large dictionary a momentum update is set to update the key in the encoder by backpropagation.

3.2.3. Multiple Instance Learning

The goal is to develop a MIL approach for the multi-label classification of lung cancer based on histopathology lung images. The dataset consists of WSIs of lung tissue samples, where each WSI represents a bag containing multiple image patches or instances. The patches can be classified into four classes: SCLC, LUAD, LUSC, or normal cells. The objective is to train a model that can accurately predict the presence of these cancer types within each WSI, which allows for the identification and classification of different cancer types in a single image. The MIL framework is suitable for this task as the exact location and quantity of cancerous cells within a WSI may vary. By using the MIL approach, completely avoids the need for a pathologist to assign instance-level labels like in the fully-supervised or semi-supervised learning approach and trains a weak model using only the WSI-level labels.

In the MIL problem, instead of a single instance like in typical classification problems such as the one presented using the ImageNet dataset, there is a bag of instances that represent one unique label. The instances should not depend on each other and their order within the bag should not be considered as significant.

These two strong definitions imply that the model must be permutation-invariant. Therefore, the permutation-invariant bag probability is computed using a scoring function for a set of instances that is a symmetric function (Zaheer et al., 2018). The score function is used to compute the bag probability and a permutation-invariant function referred to as MIL pooling ensures that this score function is a symmetric function by using commonly MIL pooling the max or mean operators.

The choice of these functions determines two different approaches to modelling the label probability as described in Section 2.3, the instance-level and embedding-level aggregation. The second aggregation is selected as proposed by Ilse et al. (2018) as the ground truth of the instances is not known, therefore, the first approach will be potentially trained insufficiently, and the prediction will be probably lower than the second approach. Moreover, each WSI has a different number of patches, in the instance-level aggregation this would result in inconsistent matrix sizes for the attention score. Instead in the embedding aggregation, the matrix is always fixed to an embedded score (c,f), being c the number of classes and f the number of features coming from the feature extractor as shown in Figure 3.C.

Zaheer et al. (2018) propose a new strategy regarding the MIL pooling layer. Instead of using the more typical max or average pooling layer, they decided to introduce an attention-based MIL pooling layer. The main difference is that the old-fashion pooling layers are predefined and non-trainable. Alternatively, with this strategy, an adaptive and flexible trainable attention pooling layer could benefit from adjusting its parameters during training to the specific task. This attention mechanism works using a weighted average of instances where the weights are trained using a neural network. As an activation function, the hyperbolic tangent (tanh) includes both negative and positive values for proper gradient flow. Together with this tanh a gated attention mechanism is implemented introducing a learnable non-linearity.

The result of this attention-based MIL pooling layer is passed to the classifier (a linear fully-connected layer) that outputs the model prediction for each class. As presented above, in this work the idea is to be able to classify if necessary more than one class at a time, as in the real case scenario when more than one cancer can be present in the same WSI. To achieve that, the final prediction is transformed into a probability using the sigmoid function and applied individually for each class, instead of the typical softmax applied in multi-class problems. If the probability is higher than 0.5 for a given class the model prediction is positive and negative otherwise.

3.3. Experimental set-up

3.3.1. K-fold cross-validation

For the training step of the MIL model, the WSIs coming from both hospitals were divided into train and validation following k-fold cross-validation. The training and validation sets are carefully split to avoid having images from the same patient in the different sets. The goal is to prove the robustness of the model to the selected training data.

The training data is divided into k (k=5) groups In each training iteration, the data from k-1 groups are utilized to train the CNN, while the remaining group is used for validation. This division is performed at the patient level to ensure that images are not shared between the training and validation partitions. Subsequently, the CNN is evaluated on the test partition, and the average and standard deviation of the k models are reported.

3.3.2. Hyperparamertes

Self-supervised pre-training: The self-supervised model is trained using all the patches available from the AOEC dataset (2,950,251 images). With this amount of data, a single experiment takes around 100 hours using an Nvidia A100 80GB. An initial learning rate (LR) of 0.03 and a MultiStepLR scheduler that decreases the LR in epochs 3, 6 and 11 by a gamma factor of 0.5. As the optimizer, Adam from Pytorch was chosen and a batch size of 256. In both cases, the last fully connected layer has the same size resulting in a 128-feature vector as output. The temperature, τ is set to 0.07 and the queue to 32,768. For the data augmentation, strong transformations are applied (Chen et al., 2020a) using the Albuminations Python library (Buslaev et al., 2020). With a probability of 0.5, the following transformations are chosen: random resize crop, vertical and horizontal flips, a random rotation of 90º, hue value saturation, colour jitter, elastic transformations, grid distortion, blurring, optical distortion, histogram equalisation and with a probability of 0.2 converting the images to grey.

Weakly-supervised training: To train the MIL model, the 128-feature vector is loaded for all patches per WSI and only the MIL pooling attention-based and the classifier are trained. Training a model last around five hours for the five models using the 5-fold cross-validation strategy using an Nvidia A100 80GB. Therefore, a grid search is performed to find the best hyperparameters to train the model for our specific structure. As shown in Table 3 several experiments are conducted combining different learning rates, optimizers, schedulers strategies, and two different loss functions specific to work with the multi-label paradigm and to address the class imbalance of our dataset as shown in Table 2. The BCEWithLogitsLoss with weights directly loaded from the Pythroch framework (Paszke et al., 2019) and the focal loss as shown in Equation 2 implemented from the work of (Lin et al., 2017).

$$\text{FocalLoss}(p_t) = -\alpha_t(1 - p_t)^{\gamma} \log(p_t) \qquad (2)$$

p_t represents the predicted probability of the true

class, and it is obtained by applying a sigmoid activation function to the logits. By multiplying the negative logarithm of p_t with the balancing factor α_t and the focusing factor γ, the focal loss penalizes misclassified examples more strongly in the class with fewer examples and thus helps to address the class imbalance and improves the training of models in the presence of difficult examples.

Table 3: Hyperparameters grid search to find the optimal values for the training of the model. LR: Learning Rate, BCELoss: Binary Cross Entropy with Logits Loss, SGD: Stochastic Gradient Descent, RMSProp: Root Mean Squared Propagation.

Batch size	Criterion	LR	Scheduler	Optimizer
256 / 512	BCELoss / Focal Loss	0.01 - 0.0001	MultiStepLR / CosineAnnealingLR	Adam / AdamW SGD / RMSProp

3.4. Evaluation

3.4.1. Self-supervised model

The performance of the self-supervised models is evaluated from two different points of view. Qualitatively, an approach to have an idea of the performance of the model before implementing it in the downstream classification task. The main idea is to interpret if the model is learning concepts from the patches. An expert pathologist selected patches from 10 different WSIs that contains cells, glands or stroma. From the feature vectors of these patches extracted from the already-trained self-supervised feature extractor. Afterwards, a Principal Component Analysis (PCA) is performed to reduce the number of features to 20 components and from there compute a dimension reduction using (UMAP) (McInnes et al., 2018) to finally reduce the 128 feature vector from the self-supervised model to 2 representative dimensions for each patch and plot the result. The goal is to evaluate if the model is really extracting separable feature vectors and learning that patches with glands, stroma or cells are different. Quantitatively, the results using the different self-supervised models are compared in the lung cancer classification downstream task with the pre-trained models on ImageNet explained in the following section.

3.4.2. Classification task

The different prevalence of the lung cancer subtypes leads to an imbalanced dataset. The major cancer type is LUAD being the class with more patients in the dataset as shown in Table 2. More important, we are working on a multi-label scenario, and different from a multi-class problem more than one class could be positive for the same WSI. These facts make it not very convenient to use metrics such as accuracy. To show the results obtain in the different models presented in this work and further comparison, the receiver operating characteristic (ROC) curve is computed for each class and the average-micro ROC curve as a global metric of the model. Together with the ROC, the AUC is computed for each class and on micro average. The idea is to understand the balance between true positive rate (TPR) and false positive rate (FPR) at different thresholds with the ROC curve. The AUC provides a global metric of the performance of the model individually for each class and globally with the average micro-AUC. The f1-score is also evaluated which gives a global idea of the precision/recall metrics of the model (Wu and Zhou, 2017). For evaluation of the model in real-time the precision/recall curve was drawn for each epoch together with the ROC curve and the AUC for training and validation. For the final evaluation of the model, all the metrics are calculated on the test set with WSIs from different patients never seen in the training phase.

3.5. Qualitative evaluation

In the MIL training, as presented above, a multi-head attention layer is used for the MIL pooling. This layer provides an attention score for all the extracted patches in a WSI per class. The idea is to provide a visualisation tool that overlaps the attention of the score for a given class in the WSI. Through this process, the pathologist would be able to understand better in which regions the model is predicting a given class. The results are presented in the form of heatmaps (Lu et al., 2021) to interpret that the model is correctly looking at where the malignancy is present on the histopathology slide of the lung.

This tool is very important to show at inference time if is really learning from the areas of interest on the WSI and not from arbitrary patches on the slides with no pathological meaning. Moreover, in clinical practice is a potential tool to help the pathologist not only receive a prediction from the model but also present in the form of a report the prediction of the model together with these heatmaps. This will potentially reduce the time necessary for a pathologist to analyse a slide presenting in the result regions on the WSI where the cancer is likely located on the slide.

4. Results

This section presents the results obtained in the 4-class classification task among SCLC, LUAD and LUSC as cancer subtypes and normal healthy tissue. The section is divided into several subsections to present different studies performed on this work:

- Comparison between the results obtained in the model trained using weak labels from a pathologist and the same model trained on weak labels coming directly from the reports using SKET.
- Evaluation of the performance between using a self-supervised model as a feature extractor or a

Table 4: Results of the lung cancer subtype classification using the model trained on the Azienda Ospedaliera per l'Emergenza Cannizzaro Catania (AOEC) dataset. The table shows the metrics tested on AOEC and AOEC and RUMC datasets using the labels generated by Semantic Knowl-edge Extractor Tool (SKET) and the ground truth provided by an expert pathologist. SCLC: small-cell lung carcinoma, LUAD: Non-small-cell lung adenocarcinoma, LUSC: Non-small-cell lung squamous cell carcinoma, AUC: Area under the curve.

Labels	**AUC SCLC**	**AUC LUAD**	**AUC LUSC**	**AUC Normal**	**micro-AUC**	**weighted f1-score**
Test on AOEC:						
SKET	0.8805 ± 0.0153	0.7780 ± 0.040	0.8331± 0.0422	0.5825 ± 0.0855	0.8037 ± 0.0282	**0.6250 ± 0.0308**
Ground truth	0.83333 ± 0.0316	0.7744 ± 0.1192	0.8275 ± 0.0430	0.6808 ± 0.0903	0.8024 ± 0.0450	0.5945 ± 0.0749
Test on AOEC and RUMC:						
SKET	0.7860 ± 0.0446	0.6507 ± 0.0384	0.7682 ± 0.0273	0.6266 ± 0.054	0.6440 ± 0.0671	0.5123 ± 0.0019
Ground truth	0.7779± 0.0540	0.67985 ± 0.0352	0.7574± 0.0180	0.7234 ± 0.0718	0.6604 ± 0.0493	0.5068 ± 0.0342

Table 5: Results of the lung cancer subtype classification using the model trained on the Azienda Ospedaliera per l'Emergenza Cannizzaro Catania (AOEC) dataset. The table shows the metrics tested on AOEC and AOEC and Radboud University Medical Centre (RUMC) datasets and compares the performance of the self-supervised pre-training model and the model pre-trained on ImageNet. SCLC: small-cell lung carcinoma, LUAD: Non-small-cell lung adenocarcinoma, LUSC: Non-small-cell lung squamous cell carcinoma, AUC: Area under the curve.

Pre-training	**AUC SCLC**	**AUC LUAD**	**AUC LUSC**	**AUC Normal**	**micro-AUC**	**weighted f1-score**
Test on AOEC:						
ImageNet	0.7766 ± 0.0369	0.7176 ± 0.0642	0.8093 ± 0.0547	0.5500 ± 0.0410	0.7264 ± 0.0305	0.5175 ± 0.0627
Self-supervised (AOEC)	0.83333 ± 0.0316	0.7744 ± 0.1192	0.8275 ± 0.0430	0.6808 ± 0.0903	0.8024 ± 0.0450	**0.5945 ± 0.0749**
Test on AOEC and RUMC:						
ImageNet	0.7506 ± 0.065	0.6812 ± 0.0357	0.8007 ± 0.0462	0.7242 ± 0.0570	0.6603 ± 0.0469	0.5314 ± 0.0389
Self-supervised (AOEC)	0.7779± 0.0540	0.67985 ± 0.0352	0.7574± 0.0180	0.7234 ± 0.0718	0.6604 ± 0.0493	0.5068 ± 0.0342

model trained on using the weights directly from ImageNet. In this section, the model trained on AOEC is compared with the model trained using data from 2 different hospitals (AOEC and RUMC). The objective is to study the importance of heterogeneity and the amount of data in the training phase.

- Analysis of the generalisation capabilities of the trained models by conducting a test study on the data from the TCGA public dataset.
- Presentation of heatmaps to show where the model is looking on the WSI and compare where a pathologist finds the lung cancer pathology.

4.1. Manual versus Automatic labels

The results of the model were trained only on data from AOEC with the main goal of comparing the different outcomes obtained using manual or automatic labels. In table 4 are illustrated the results of both strategies tested on the unseen test set during training. There are two test sets, one composed of only AOEC data and the other composed of data from AOEC and RUMC to study the performance in a different dataset.

4.2. Self-supervised validation

To evaluate the self-supervision performance two different results are presented. Quantitatively, the effectiveness of the weakly-supervised classification model is compared through two different scenarios. Pre-training the model with the self-supervision model and simply loading the pre-trained model from ImageNet. Qualitatively, UMAPs of the self-supervision model and the model pre-train on ImageNet to visualize its capabilities to cluster different types of patches.

4.2.1. Self-supervised vs ImageNet pre-training

Table 5 presents the results of the model trained on data from AOEC and pre-trained using self-supervised learning in comparison with the same model with the same fine-tuning using the pre-trained model from ImageNet. The model is tested on both, the AOEC dataset and AOEC and RUMC datasets.

Table 6 illustrates the results of the model trained on data from AOEC and RUMC and pre-trained using self-supervised learning in comparison with the same model with the same fine-tuning using the pre-trained model from ImageNet. The model is tested on AOEC and RUMC datasets together.

Figure 5 shows the ROC curves and AUC for the models pre-trained using the self-supervised train model on AOEC (top image) and the models presented on ImageNet (bottom image). On the left side are shown the results of the MIL models trained and tested on AOEC and AOEC and RUMC, respectively. On the right side are presented the models trained on data from AOEC and RUMC and tested on TCGA.

Table 6: Results of the lung cancer subtype classification using the model trained on the Azienda Ospedaliera per l'Emergenza Cannizzaro Catania (AOEC) and Radboud University Medical Centre (RUMC) datasets. The table shows the metrics tested on AOEC and RUMC and compares the performance of the self-supervised pre-training model and the model pre-trained on ImageNet. SCLC: small-cell lung carcinoma, LUAD: Non-small-cell lung adenocarcinoma, LUSC: Non-small-cell lung squamous cell carcinoma, AUC: Area under the curve.

Pre-training	**AUC SCLC**	**AUC LUAD**	**AUC LUSC**	**AUC Normal**	**micro-AUC**	**weighted f1-score**
Test on AOEC and RUMC:						
ImageNet	0.8784± 0.096	0.7497 ± 0.0268	0.8764 ± 0.0247	0.8446 ± 0.0071	0.8596 ± 0.0143	0.6380 ± 0.0148
Self-supervised (AOEC + RUMC)	0.8825 ± 0.0712	0.7457 ± 0.0267	0.8428 ± 0.0171	0.8468 ± 0.0130	0.8558 ± 0.0051	**0.6537 ± 0.0237**

Table 7: Results of the lung cancer subtype classification of the two both models, trained in the Azienda Ospedaliera per l'Emergenza Cannizzaro Catania (AOEC) dataset and the model trained on the AOEC and Radboud University Medical Centre (RUMC) datasets and tested on The Cancer Genome Atlas (TCGA) public dataset. SCLC: small-cell lung carcinoma, LUAD: Non-small-cell lung adenocarcinoma, LUSC: Non-small-cell lung squamous cell carcinoma, AUC: Area under the curve.

Pre-training	**AUC SCLC**	**AUC LUAD**	**AUC LUSC**	**AUC Normal**	**micro-AUC**	**weighted f1-score**
Test on TCGA:						
Train on AOEC:						
ImageNet	1.0 ± 0.0	0.8754± 0.0081	0.8639 ± 0.0191	1.0 ± 0.0	0.9215 ± 0.0323	0.7212 ± 0.073
Self-supervised (AOEC)	1.0 ± 0.0	0.8464 ± 0.0290	0.8735 ± 0.0370	1.0 ± 0.0	0.8762 ± 0.0205	0.6688 ± 0.0143
Train on AOEC and RUMC:						
ImageNet	1.0 ± 0.0	0.8861 ± 0.0178	0.8875 ± 0.0168	1.0 ± 0.0	0.9448 ± 0.0078	**0.7737 ± 0.0259**
Self-supervised (AOEC + RUMC)	1.0 ± 0.0	0.8818 ± 0.0163	0.8856 ± 0.0179	1.0 ± 0.0	0.9433 ± 0.0198	**0.7726 ± 0.0438**

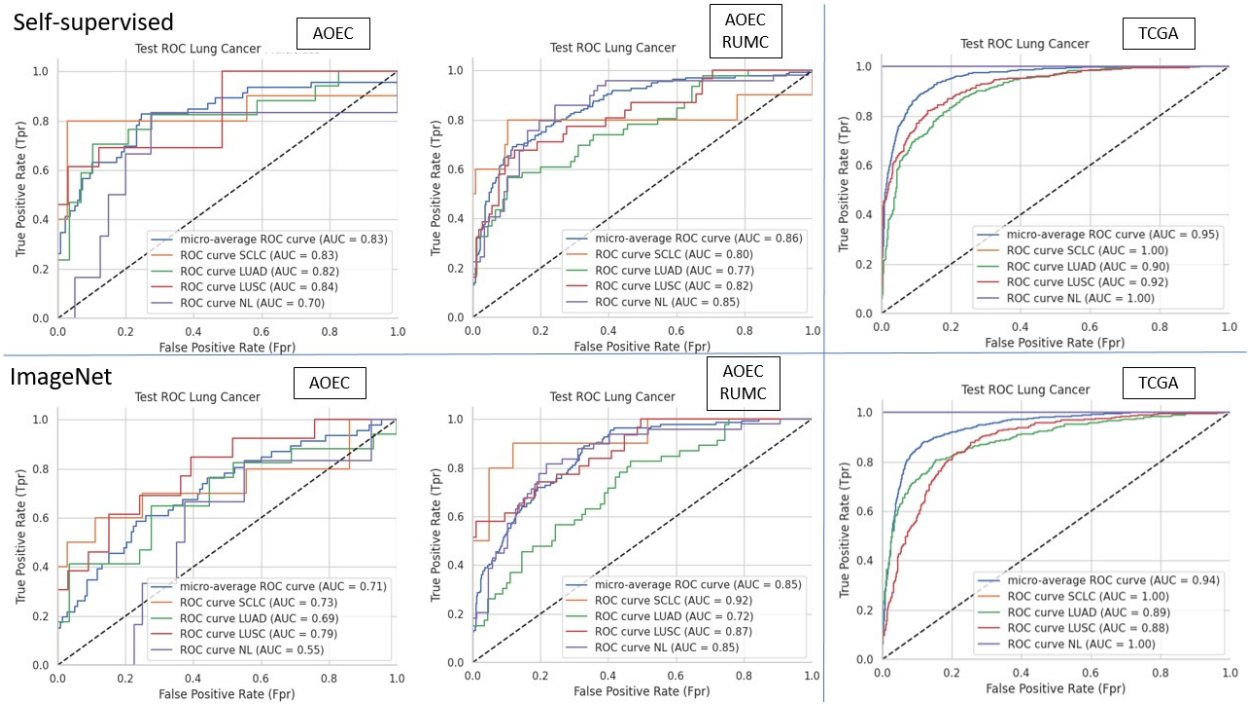


Figure 5: Receiver operating characteristic (ROC) curves and area under the curve (AUC) in the 4-class classification task among three cancer subtypes, small-cell lung carcinoma (SCLC), non-small-cell lung adenocarcinoma (LUAD) and non-small-cell lung squamous cell carcinoma LUSC and normal tissue (NL). The top of the image presents the models pre-trained using the self-supervised model trained on Azienda Ospedaliera per l'Emergenza Cannizzaro Catania (AOEC), while the bottom shows the models pre-trained on ImageNet. The left area shows the models train on AOEC and AOEC and Radboud University Medical Centre (RUMC), respectively. On the right side are illustrated the results of the model trained on AOEC and RUMC and testing them on The Cancer Genome Atlas (TCGA) public dataset.

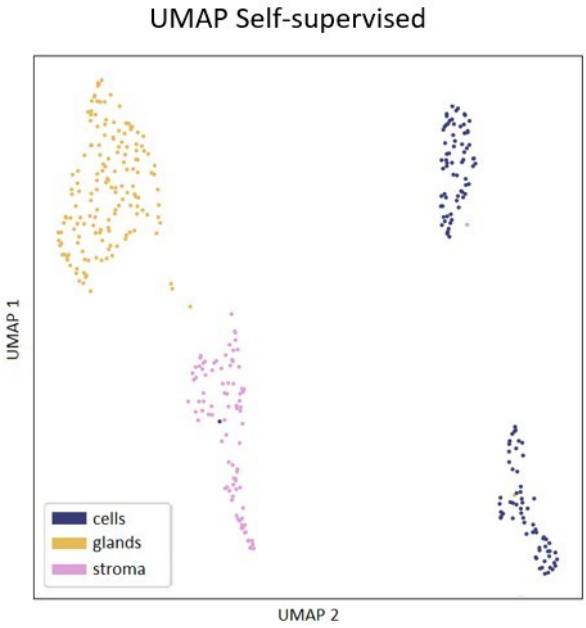


Figure 6: Uniform Manifold Approximation and Projection (UMAP) for dimension reduction computed on 384 patches (135 cells, 158 glands and 91 stroma), selected by an expert pathologist, using the self-supervised trained model.

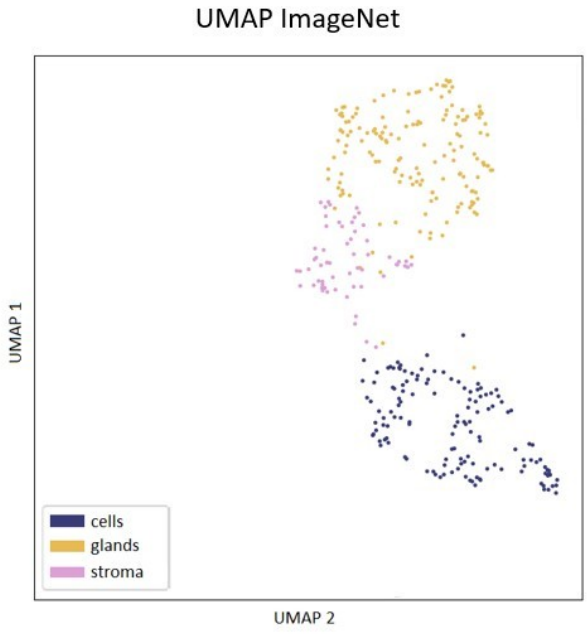


Figure 7: Uniform Manifold Approximation and Projection (UMAP) for dimension reduction computed on 384 patches (135 cells, 158 glands and 91 stroma), selected by an expert pathologist, using the pre-trained model from ImageNet.

4.2.2. Clustering proficiency

For the self-supervised learning as mentioned above the best model is chosen by the results obtained in the lung cancer subtype classification task. Nevertheless, in this section, UMAPs are computed from the features extracted from the self-supervised model and the model pre-train on ImageNet. The UMAP is computed on 384 patches (135 cells, 158 glands and 91 stroma) selected by an expert pathologist. The UMAP for the self-supervised model is shown in Figure 6 and the UMAP for the pre-trained model from ImageNet is illustrated in Figure 7.

4.2.3. Test on the public TCGA dataset

The results of the models tested on the public TCGA dataset are illustrated in Table 7. The model trained only on data from AOEC and the model trained using the AOEC and RUMC datasets are presented using both, self-supervised pre-training and the models pre-trained on ImageNet.

4.2.4. Qualitative evaluation: Heatmaps

For qualitative evaluation of the MIL model performance, a heatmap is illustrated in Figure 8 computed on the model with the best performance. The model is pre-trained using self-supervised learning and trained on the AOEC and RUMC datasets. Three more heatmaps are presented in the Annex to illustrate a wider representation of the study with three more examples, one for each cancer subtype.

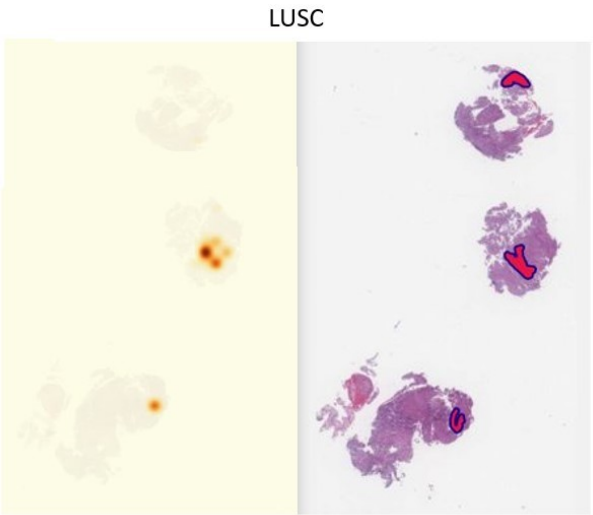


Figure 8: Heatmap computed using the final trained model (left side) compared with the annotations from an expert pathologist (right side) of a whole slide image (WSI) diagnosed with non-small-cell lung squamous cell carcinoma (LUSC).

5. Discussion

The main idea of this thesis was to develop a full pipeline to build an algorithm able to classify, using lung WSIs, between the three most prevalent cancer types, LUAD, LUSC, SCLC and normal tissue. Because of the size of the WSIs, composed of billions of pixels, the strategy adopted in this work, is first preprocessing all the datasets and splitting the WSIs into non-overlapping patches of size 256x256 at a 10x magnification level. Afterwards, the model is pre-trained with self-supervised learning using Momentum Contrastive Learning (MoCo). Finally, a MIL model is trained to perform the classification task using weakly-supervised learning. The performance of the model is evaluated in four different scenarios to understand the different strengths and limitations of the strategy chosen and the training process of the model as presented in the different results sections.

The performance of the model is evaluated while training using weak labels coming from two different sources. Using automatically made labels by SKET (Marchesin et al., 2022; Marini et al., 2022) and manually labels made by an expert pathologist. As shown in Table 4 the performance of both models are very similar. Trained and tested on the AOEC dataset the model using automatic labels achieved a weighted f1-score of 0.6250 ± 0.0308 in comparison with 0.5945 ± 0.0749 obtained using manual labels. Regarding the AUC both models obtained very similar average micro-AUC. While both models have similar values in predicting LUAD and LUSC, the model using automatic labels outperformed the network with manual labels when predicting SCLC but the model using ground truth labels outperformed the rival when predicting normal tissue samples. This is an important finding that supports the works of Marchesin et al. (2022); Marini et al. (2022) that confirms the possibility to train deep learning models with weakly-supervised learning using automatic labels directly extracted from the reports unleashing the potential of histopathological lung datasets without labels for predicting lung cancer.

The performance of the model can be influenced by the amount of heterogeneous data. This is a very common scenario in computational pathology where big differences are found in the WSIs from the presence of different H&E stains to differences in the characteristics of the scanners. Building on the findings in Table 4 is appreciable that models trained on AOEC obtain better results when tested on the same training dataset than when tested combining WSIs from an Italian and Dutch hospital. To improve the model on top of these observations, a new MIL model is trained, in this case, using data from both hospitals. As illustrated in Table 6 the performance when testing on both hospitals improved from a f1-score of 0.5068 ± 0.0342 when trained on AOEC to a f1-score of 0.6537 ± 0.0237 when trained in both datasets. The model effectively improved the performance considerably because it was trained with more heterogeneous data combining both datasets for training and also potentially because of the increase in the total number of WSIs used to train the model.

For the validation of the self-supervision learning strategy, two different approaches were adopted. Evaluate quantitatively the performance of the model against the same model pre-trained on ImageNet. Qualitatively compare the UMAPs of the same models mentioned above. When training in the AOEC dataset the self-supervised model outperformed the model pre-trained on ImageNet with a gap of almost 0.1 on the f1-score. Moreover, when both models were trained using both datasets (AOEC and RUMC) the self-supervision model surpass the pre-trained model with an f1-score of 0.6537 ± 0.0237 and 0.6380 ± 0.0148, respectively. These observations support the fact that a model learns more accurately when trained on high-level features representative of the dataset, in this case, histopathological lung data, than on images from a model pre-trained on ImageNet (dogs, vehicles, etc).

In Figure 5 the ROC curves and AUC results are presented between the three models pre-trained with self-supervision using the AOEC dataset and the same three models pre-trained on ImageNet. On the left side, when the MIL models are trained on data from the Catania hospital there is a big improvement as a consequence of the self-supervised learning strategy. However, when the MIL models are trained on AOEC and RUMC datasets the improvement is not that noticeable. Possibly, the fact that the self-supervised model is only trained on AOEC data could be the main reason. Therefore, the effect of self-supervision will be more significant if the training is done with more data.

Qualitatively, the UMAP is plotted for 384 patches composed of cells glands are stroma. As shown in Figure 6 the self-supervised model is able to almost perfect cluster the three types of patches into clear separable regions with as exception of some outliers. Contrarily, in the UMAP of the pre-train model on Image-Net (Figure 7) is possible to appreciate that the model does not separate perfectly the different clusters with some overlaps on the regions between cells and stroma and glands and stroma.

To evaluate the generalisation capabilities of the trained models, they were tested on the public TCGA dataset. This dataset is composed of 1,036 WSIs of LUAD and LUSC collected from 5 different centres in the USA. The best models are both trained on the AOEC and RUMC datasets using self-supervision and ImageNet pre-training strategies with a weighted f1-score of 0.7726 ± 0.0438 and 0.7737 ± 0.0259, respectively. These results point out the good generalisation capabilities of both models performing good predictions on an unseen dataset from another country. As discussed in the previous paragraph, in Table 7, is clear that both

models trained on data from AOEC and RUMC surpass both models trained only on the AOEC dataset. This also supports the fact that a model trained on more heterogeneous data is able to perform predictions more accurately than its counterpart.

For the qualitative evaluation of the networks, a tool was developed to elaborate heatmaps. These heatmaps are computed from the attention scores coming from the multi-head attention of the MIL model for each class of a given WSI. On these heatmaps, only the attention scores of the ground truth class are computed and compared with the manual annotations from an expert pathologist. As shown in Figure 8 the model is accurately giving high scores for the LUSC class to the patches that are in the region similar to the manual annotations of the pathologist. Of the three areas where the pathologist indicates that there is the presence of LUSC two are localized for the model a no false positives are given. Additionally, in the Annex is possible to find 3 more examples. The only case where the model is giving more false positives is in the SCLC probably because of the minor number of examples present in the dataset. Nevertheless, also two out of three regions annotated by the pathologist are localized by the model.

Fully-supervised learning is usually the best approach in terms of performance for training models to classify lung histopathological data as presented in Section 2 with the work of Coudray et al. (2018); Kanavati et al. (2021); Yang et al. (2021). The major drawbacks of this approach are the need for pixel-level annotations for training which is a very time-consuming task and that this approach does not simulate the real scenario where one label (or more if more than one malignancy is present) is reported per WSI. Nevertheless, our model with weakly-supervised learning has similar results than the fully-supervised model presented by Yu et al. (2020) and surpasses the work of Le Page et al. (2021). Moreover, the self-supervised model achieves this good performance in a 4-class classification task while the two papers mentioned above only present a binary classification. Among the weakly-supervised strategies, our model surpasses the performance of and Lu et al. (2021) with an AUC of 0.902 ± 0.016 and obtains similar results than Chen et al. (2021). Both works used transfer learning as a pre-train strategy in comparison with the self-supervision presented in this work.

Recent findings on the computer vision field with the work of Caron et al. (2021) with DINO, and more specifically for histopathological images, Chen et al. (2022) with HIPT, shown the potential of self-supervised learning to improve the prediction of ML models. They show that using Vision Transformers (ViT) is possible to obtain better feature representation of the images than using the architectures proposed by Chen et al. (2020b) with MoCo v2 and Chen et al. (2020a) with SimCLR using CNNs. The idea behind HIPT is to use a scaling strategy in two stages using two consecutive ViT. First patching pre-training using patches of 256x256 and on top of this another ViT that performs a region pre-training of size 4,096x4,096 using the features coming from the first stage. Finally, these regions are used as feature extractors to feed a MIL model that performs the downstream classification task. The only drawback is that training HIPT, Chen et al. (2022) uses a dataset with 10,678 WSIs, with a total of 433,779 regions of 4,096x4,096 pixels that take 7.7 TB of space. These specifications need powerful GPUs to be able to handle the training size.

As proven in this work, the training of MIL models potentially takes advantage of self-supervised pre-training. Chen et al. (2022) using HIPT obtain an AUC of 0.952 ± 0.021 in the binary classification between LUAD and LUSC using 1,008 WSIs from the TCGA dataset. In our case for the 4-class classification presented in this work, the MIL model obtains an AUC of 0.9448 ± 0.0078 but AUCs of 0.8818 ± 0.0163 and 0.8856 ± 0.0179 in LUAD and LUSC, respectively. HIPT obtains better performance by using the scaling ViT strategy.

Improving the performance of CAD systems on computational pathology and creating powerful tools for clinical diagnosis to alleviate pathologists' workload is still an open challenge. Perhaps, it would be very interesting to combine the power of HIPT as a feature extractor and the findings of this paper, such as using SKET to have weak labels for more WSIs and training MIL models to classy, for example in the lung cancer scenario, not only LUAD and LUSC but more lung subtypes as presented in this paper mimicking real clinical scenario. Regarding the recreation of a clinical real-life scenario, include a multi-label strategy for training as more than one cancer subtype can be present on the same WSI as it is presented in this work. Additionally, combining more datasets from different hospitals and public datasets to train a more robust MIL model using more heterogeneous data.

6. Conclusions

There are different learning approaches to train the histopathology images, fully-supervised achieves the best results but not simulates the real scenario, in which, usually only a singular label per WSI is provided (or more than one, multi-label scenario as presented in this work). Weakly-supervised learning tries to address this issue by training a model with only WSI-level annotations. The algorithm proposed in this work is first pre-trained using self-supervised learning to extract high-level features representative of the dataset. Afterwards, using weakly-supervised learning, a MIL model is trained on the 4-class classification task to predict among three cancer types, SCLC, LUAD and LUSC, and normal tissue. It is demonstrated through the different experiments conducted on this work the

following findings. First, the ability of the model to be trained using automatic labels extracted directly from the pathologist reports unleashes the potential of using unlabeled datasets. Second, a model trained on a more heterogeneous dataset and with a larger number of WSIs would potentially increase the performance. Moreover, it is proven that training with a more heterogeneous dataset also improves the generalisation capabilities of the model making predictions in unseen data from external datasets. Third, self-supervised learning is able to elaborate high-level features more representative of the dataset, used later on for training, than directly using a pre-train model on Imagenet. This is shown both, quantitatively through the lung subtype classification downstream task and qualitatively through the UMAP representation of labelled patches. Finally, a tool is provided in order to interpret better where the model is actually looking when making a prediction to a given class. These heatmaps together with the prediction of the model would be a powerful tool that provided in clinical scenarios to the pathologist could potentially reduce the time performed to analyse a new WSI. Developing CAD systems to help pathologists make a diagnosis will be beneficial in a world where the number of pathologists is not increasing linearly with the number of biopsies and resections performed. All code was implemented in Python using PyTorch as the primary DL library. The repository includes the full pipeline from processing the WSIs to the training and evaluation of the models, and is available at **https://github.com/lluisb3/histo_lung**

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825292 (ExaMode, http://www.examode.eu/).

References

Abels, E., Pantanowitz, L., Aeffner, F., Zarella, M.D., Laak, J., Bui, M.M., Vemuri, V.N., Parwani, A.V., Gibbs, J., Agosto-Arroyo, E., et al., 2019. Computational pathology definitions, best practices, and recommendations for regulatory guidance: A white paper from the digital pathology association. The Journal of Pathology 249, 286–294. doi:10.1002/path.5331.

Albertina, B., Watson, M., Holback, C., Jarosz, R., Kirk, S., Lee, Y., Rieger-Christ, K., Lemmerman, J., 2016. The cancer genome atlas lung adenocarcinoma collection (tcga-luad) (version 4) [data set]. The Cancer ImagingF Archive. https://doi.org/10.7937/K9/TCIA.2016.JGNIHEP5.

American Cancer Society, . Cancer statistics center. http://cancerstatisticscenter.cancer.org. Accessed April 12, 2023.

Boykov, Y., Kolmogorov, V., 2004. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Transactions on Pattern Analysis and Machine Intelligence 26, 1124–1137. doi:10.1109/tpami.2004.60.

Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A., 2020. Albumentations: Fast and flexible image augmentations. Information 11. URL: https://www.mdpi.com/2078-2489/11/2/125, doi:10.3390/info11020125.

Canny, J., 1986. A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-8, 679–698. doi:10.1109/tpami.1986.4767851.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 9650–9660.

Chen, C.L., Chen, C.C., Yu, W.H., Chen, S.H., Chang, Y.C., Hsu, T.I., Hsiao, M., Yeh, C.Y., Chen, C.Y., 2021. An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. Nature Communications 12. doi:10.1038/s41467-021-21467-y.

Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F., 2022. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16144–16155.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020a. A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR. pp. 1597–1607.

Chen, X., Fan, H., Girshick, R., He, K., 2020b. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 .

Cheplygina, V., de Bruijne, M., Pluim, J.P., 2019. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical Image Analysis 54, 280–296. doi:10.1016/j.media.2019.03.009.

Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., Tsirigos, A., 2018. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature Medicine 24, 1559–1567. doi:10.1038/s41591-018-0177-5.

Dehaene, O., Camara, A., Moindrot, O., de Lavergne, A., Courtiol, P., 2020. Self-supervision closes the gap between weak and strong supervision in histology. arXiv preprint arXiv:2012.03583 .

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255.

Felzenszwalb, P.F., Huttenlocher, D.P., 2004. Efficient graph-based image segmentation. International Journal of Computer Vision 59, 167–181. doi:10.1023/b:visi.0000022288.19776.77.

Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D.M., Piñeros, M., Znaor, A., Bray, F., 2021. Cancer statistics for the year 2020: An overview. International Journal of Cancer 149, 778–789. doi:10.1002/ijc.33588.

Ferlay, J., Ervik, M., Lam, F., Colombet, M., Mery, L., Piñeros, M., Znaor, A., Soerjomataram, I., Bray, F., 2020. Cancer today. URL: https://gco.iarc.fr/today/.

Fischer, A.H., Jacobson, K.A., Rose, J., Zeller, R., 2008. Hematoxylin and eosin staining of tissue and cell sections. Cold Spring Harbor Protocols 2008. doi:10.1101/pdb.prot4986.

Goldstraw, P., Ball, D., Jett, J.R., Le Chevalier, T., Lim, E., Nicholson, A.G., Shepherd, F.A., 2011. Non-small-cell lung cancer. The Lancet 378, 1727–1740.

Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06) doi:10.1109/cvpr.2006.100.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9729–9738.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Ilse, M., Tomczak, J., Welling, M., 2018. Attention-based deep multiple instance learning, in: International conference on machine

learning, PMLR. pp. 2127–2136.

Kanavati, F., Toyokawa, G., Momosaki, S., Rambeau, M., Kozuma, Y., Shoji, F., Yamazaki, K., Takeo, S., Iizuka, O., Tsuneki, M., 2020. Weakly-supervised learning for lung carcinoma classification using deep learning. Scientific Reports 10. doi:10.1038/s41598-020-66333-x.

Kanavati, F., Toyokawa, G., Momosaki, S., Takeoka, H., Okamoto, M., Yamazaki, K., Takeo, S., Iizuka, O., Tsuneki, M., 2021. A deep learning model for the classification of indeterminate lung carcinoma in biopsy whole slide images. Scientific Reports 11. doi:10.1038/s41598-021-87644-7.

Kirk, S., Lee, Y., Kumar, P., Filippini, J., Albertina, B., Watson, M., Rieger-Christ, K., Lemmerman, J., 2016. The cancer genome atlas lung squamous cell carcinoma collection (tcga-lusc) (version 4) [data set]. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2016.TYGKKFMQ.

Kumar, V., Abbas, A.K., Aster, J.C., Perkins, J.A., 2018. Robbins basic pathology. Elsevier.

Lababede, O., Meziane, M.A., 2018. The eighth edition of tnm staging of lung cancer: Reference chart and diagrams. The Oncologist 23, 844–848. doi:10.1634/theoncologist.2017-0659.

Le Page, A.L., Ballot, E., Truntzer, C., Derangère, V., Ilie, A., Rageot, D., Bibeau, F., Ghiringhelli, F., 2021. Using a convolutional neural network for classification of squamous and non-squamous non-small cell lung cancer based on diagnostic histopathology hes images. Scientific Reports 11. doi:10.1038/s41598-021-03206-x.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988.

Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F., 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering 5, 555–570.

Marchesin, S., Giachelle, F., Marini, N., Atzori, M., Boytcheva, S., Buttafuoco, G., Ciompi, F., Di Nunzio, G.M., Fraggetta, F., Irrera, O., et al., 2022. Empowering digital pathology applications through explainable knowledge extraction tools. Journal of pathology informatics 13, 100139.

Marini, N., Marchesin, S., Otálora, S., Wodzinski, M., Caputo, A., Van Rijthoven, M., Aswolinskiy, W., Bokhorst, J.M., Podareanu, D., Petters, E., et al., 2022. Unleashing the potential of digital pathology data by training computer-aided diagnosis models without human annotations. NPJ digital medicine 5, 102.

Marini, N., Otálora, S., Podareanu, D., van Rijthoven, M., van der Laak, J., Ciompi, F., Müller, H., Atzori, M., 2021a. Multi_scale_tools: a python library to exploit multi-scale whole slide images. Frontiers in Computer Science 3, 684521.

Marini, N., Otálora, S., Müller, H., Atzori, M., 2021b. Semi-supervised training of deep convolutional neural networks with heterogeneous data and few local annotations: An experiment on prostate histopathology image classification. Medical Image Analysis 73, 102165. doi:10.1016/j.media.2021.102165.

Märkl, B., Füzesi, L., Huss, R., Bauer, S., Schaller, T., 2021. Number of pathologists in germany: comparison with european countries, usa, and canada. Virchows Archiv 478, 335–341.

McInnes, L., Healy, J., Melville, J., 2018. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 .

Muñoz-Aguirre, M., Ntasis, V.F., Rojas, S., Guigó, R., 2020. Pyhist: A histological image segmentation tool. PLOS Computational Biology 16. doi:10.1371/journal.pcbi.1008349.

Otálora, S., Marini, N., Müller, H., Atzori, M., 2021. Combining weakly and strongly supervised learning improves strong supervision in gleason pattern classification. BMC Medical Imaging 21. doi:10.1186/s12880-021-00609-0.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems 32.

Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F., 2021. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians 71, 209–249. doi:10.3322/caac.21660.

Tornillo, L., Franco, R., 2022. The role of histopathology in cancer diagnosis and prognosis. Frontiers Research Topics doi:10.3389/978-2-83250-721-6.

Travis, W.D., Brambilla, E., Noguchi, M., Nicholson, A.G., Geisinger, K.R., Yatabe, Y., Beer, D.G., Powell, C.A., Riely, G.J., Van Schil, P.E., et al., 2011. International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. Journal of Thoracic Oncology 6, 244–285. doi:10.1097/jto.0b013e318206a221.

Van Meerbeeck, J.P., Fennell, D.A., De Ruysscher, D.K., 2011. Small-cell lung cancer. The Lancet 378, 1741–1755.

Wang, X., Chen, H., Gan, C., Lin, H., Dou, Q., Tsougenis, E., Huang, Q., Cai, M., Heng, P.A., 2020. Weakly supervised deep learning for whole slide lung cancer image analysis. IEEE Transactions on Cybernetics 50, 3950–3962. doi:10.1109/tcyb.2019.2935141.

Wu, X.Z., Zhou, Z.H., 2017. A unified view of multi-label performance measures, in: international conference on machine learning, PMLR. pp. 3780–3788.

Wu, Z., Xiong, Y., Yu, S.X., Lin, D., 2018. Unsupervised feature learning via non-parametric instance discrimination, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3733–3742.

Yang, H., Chen, L., Cheng, Z., Yang, M., Wang, J., Lin, C., Wang, Y., Huang, L., Chen, Y., Peng, S., et al., 2021. Deep learning-based six-type classifier for lung cancer and mimics from histopathological whole slide images: A retrospective study. BMC Medicine 19. doi:10.1186/s12916-021-01953-2.

Yu, K.H., Wang, F., Berry, G.J., Ré, C., Altman, R.B., Snyder, M., Kohane, I.S., 2020. Classifying non-small cell lung cancer types and transcriptomic subtypes using convolutional neural networks. Journal of the American Medical Informatics Association 27, 757–769. doi:10.1093/jamia/ocz230.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R., Smola, A., 2018. Deep sets. arXiv:1703.06114.

Annex: Heatmaps

In this section, supplementary heatmaps are presented to illustrate the wider range of classes and not only one example. One example of each cancer type is presented, SCLC (Figure 9), LUAD (Figure 10) and LUSC (Figure 11).

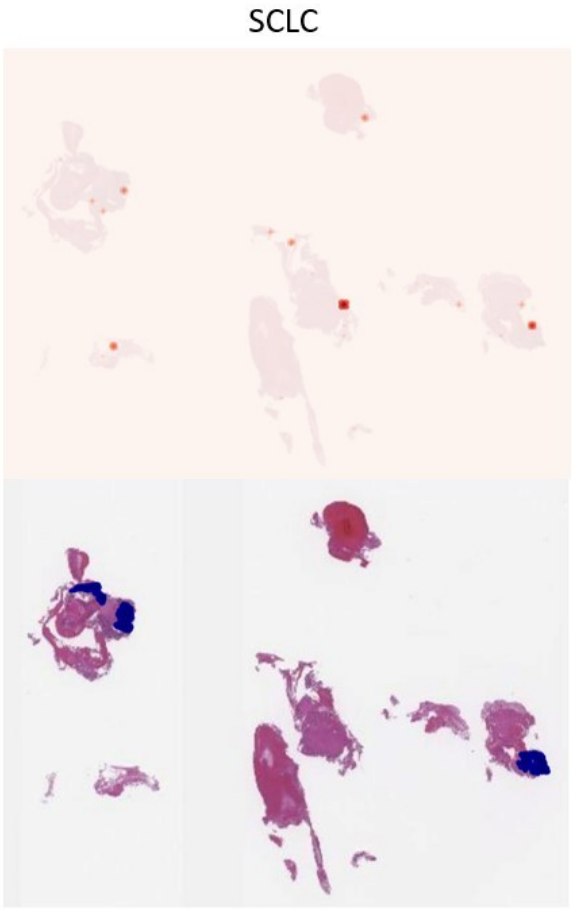


Figure 9: Heatmap computed using the final trained model (top image) compared with the annotations from an expert pathologist (bottom image) of a whole slide image (WSI) diagnosed with small-cell lung carcinoma (SCLC).

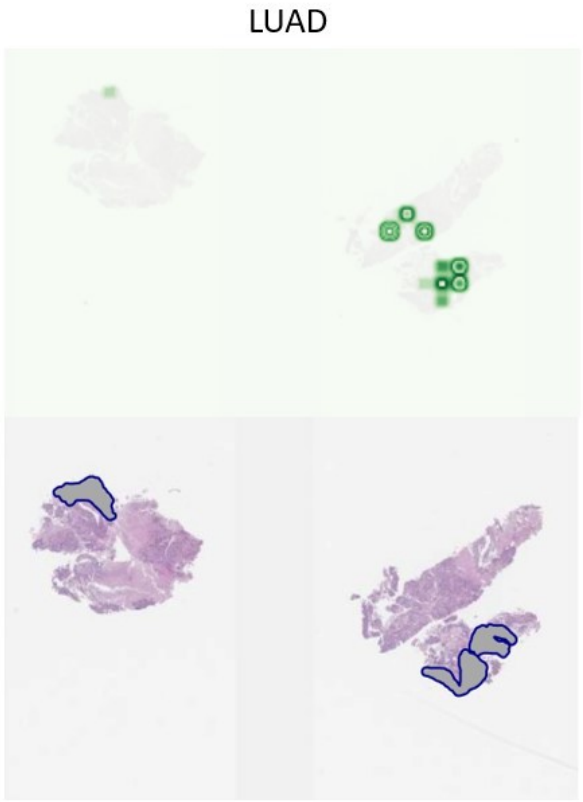


Figure 10: Heatmap computed using the final trained model (top image) compared with the annotations from an expert pathologist (bottom image) of a whole slide image (WSI) diagnosed with non-small-cell lung adenocarcinoma (LUAD).

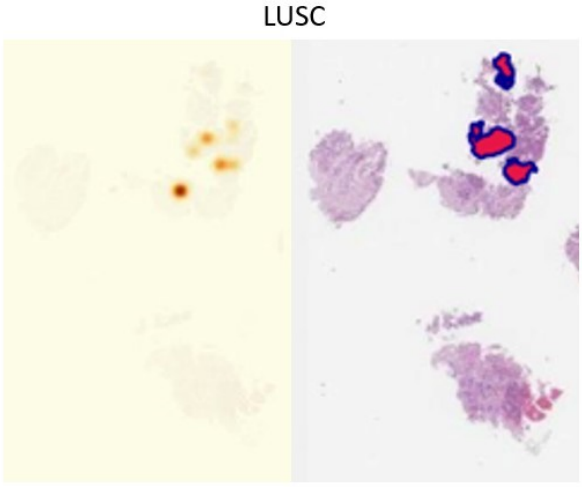
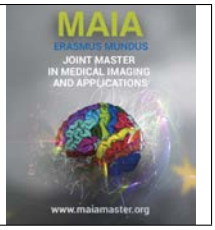


Figure 11: Heatmap computed using the final trained model (left side) compared with the annotations from an expert pathologist (right side) of a whole slide image (WSI) diagnosed with non-small-cell lung squamous cell carcinoma (LUSC).

Medical Imaging and Applications

Master Thesis, June 2023



Classification of Multiple Sclerosis Using Brain MRI and Clinical Data

Emily E. Carvajal-Camelo[a,b], Ezequiel de la Rosa[a,d], Mladen Rakić[a,c], Diana M. Sima[a]

[a] *ico**metrix**, Leuven, Belgium*
[b] *University of Girona, Girona, Spain*
[c] *Department of Electrical Engineering, KU Leuven, Leuven, Belgium,*
[d] *Department of Informatics, Technical University of Munich, Munich, Germany*

Abstract

Multiple Sclerosis (MS) is a demyelinating disease characterised by white matter lesions. These lesions can lead to physical disabilities and cognitive deficiency. The severity of the disease is commonly assessed using the Expanded Disability Status Scale score, which measures disability across various stages ranging from "non-symptomatic" to "death" and passing by different disability conditions. Accurately classifying the progression of MS is difficult due to the complex differentiation among the MS groups. The data is imbalanced, and patients' conditions can change over time as they transition from the clinically isolated syndrome stage to relapsing-remitting MS (RRMS) and later on to secondary progressive MS (SPMS). Additionally, patients who do not experience relapses are typically put as having primary progressive MS (PPMS).

In this study, we explore the potential of using the Dynamic Affine Feature Map Transform (DAFT) approach for classifying MS groups and EDSS scores. We extend the input dimensionality by concatenating multiple medical imaging sequences with a lesion mask and apply regularisation strategies such as dropout and augmentation, along with techniques derived from existing literature. Our experimental results demonstrate that the DAFT approach has higher ensemble balanced accuracy than the baseline methods that solely use imaging data or tabular information.

Keywords: multiple sclerosis, classification, DAFT, MS groups

1. Introduction

Multiple Sclerosis (MS) is a chronic neurological disease that affects approximately 2.8 million people worldwide, according to the most recent studies (Walton et al., 2020). MS causes inflammation and directly attacks the central nervous system, including the brain, spinal cord, and optic nerves (Aslam et al., 2022). Specifically, MS is characterised by the demyelination of the axons and is typically diagnosed by quantifying white matter lesions. The lesions vary in size, shape, and location. The structural damage to the central nervous system links with other MS manifestations, such as physical disability and cognitive deficits (Eijlers et al., 2018). The disease is commonly classified into different MS phenotypes, also called MS groups (Lublin et al., 2014):

- Clinically isolated syndrome (CIS) is the first episode of neurological symptoms that lasts at least 24 hours and is caused by inflammation or demyelination of the central nervous system. CIS might not meet all the criteria for MS.

- Relapsing-remitting multiple sclerosis (RRMS) is the most common type of MS, characterised by clearly defined attacks, also known as relapses or exacerbations, of new or growing neurological symptoms, followed by intervals of remission.

- Primary progressive multiple sclerosis (PPMS) is characterised by worsening of neurological functions since the onset of symptoms without early relapsing or remission.

- Secondary progressive multiple sclerosis (SPMS) is followed by an early relapse period. Patients who have more severe symptoms in their RRMS phase may experience progression to SPMS.

Diagnosing patients with MS initially places them into distinct groups. However, this categorisation can evolve over time, such as when CIS progresses into RRMS, or when the majority of RRMS patients transition into SPMS. Additionally, SPMS and primary PPMS exhibit numerous shared MRI features. Consequently, this intricate interplay among different MS subtypes makes classifying patients into distinct groups a complex task (Shoeibi et al., 2021).

In terms of physical disability, several indexes and scoring systems have been developed to assess the clinical severity and functional deficits in patients with MS. One of the most commonly used is the Expanded Disability Status Scale (EDSS), a scale that ranges from 0 to 10 with 0.5 increments. An EDSS from 1.0 to 4.5 refers to people who are able to walk without any aid. EDSS from 5.0 to 9.5 is typically characterised by the impairment to walk.

In this work, we want to use images and clinical information obtained from a sample of patients to classify MS groups and EDSS scores. To the best of our knowledge, this is the first time that these classification problems are addressed using convolutional neural networks that allow embedding clinical information in the form of tabular data along with brain MR images and lesion segmentation masks.

2. State of the art

In this section, we present a review of the related work on the classification of MS groups using imaging and tabular data. While the literature specific to MS group classification using both data forms is limited, there is information from studies focusing on other diseases with similar challenges. Notably, we found three relevant papers that explore the integration of imaging and tabular data for classification tasks in MS. The following subsections summarise the findings and discuss their potential applicability to MS group classification.

2.1. Classification of MS

2.1.1. Methods for Classification of MS using Imaging Data

Several studies have been conducted employing deep learning techniques to classify or predict MS based on imaging data as shown in Aslam et al. (2022), or Shoeibi et al. (2021).

The study of Zhang et al. (2018) for the classification between MS and no MS patients, presented a combination of the parametric Rectified Linear Unit (ReLU), PReLU, and dropout techniques, as well as data augmentation techniques. The PReLU improved the model fitting with almost no change in computational cost, while the dropout helped increase accuracy and reduce overfitting. These methods improve accuracy in models with less than 10 layers. However, the performance is degraded in deeper models.

In the same manner, the study of Wang et al. (2018) for the classification between MS and no MS patients, developed a 14-layer convolutional neural network (CNN) that improves their work using stochastic pooling, which utilises non-maximal activations within the pooling region, an improvement compared with average or max pooling. Also, dropout, batch normalisation, and data augmentation were added to overcome overfitting.

Moreover, the study of Calimeri et al. (2018) for the classification of MS groups, aimed to develop a classification method that uses structural connectivity information related to white matter networks to generate structural connectivity graphs. Tractography was used as input information for a graph-based neural network. The main issue using this method was the small dataset.

In the study of Eitel et al. (2019) for the classification between MS and no MS patients, a layer-wise relevance propagation map that enables uncovering relevant image features that CNN uses for decision-making was proposed. The map was applied to images with MS images with hyperintense lesions. They concluded that the CNN used in the paper focused on hyperintense lesions as the primary source of information, but also incorporated information from lesion locations and normal-appearing brain areas. This showed the need for explainability to retract the classification decision.

In terms of explainability, another approach was introduced by Zhang et al. (2021). They utilised Gradient-weighted Class Activation Mapping to gain insights into the decision-making process of CNNs. The study found that for SPMS cases, the CNNs highlighted frontal or temporal/parietal regions. In the case of RRMS, the focus was on frontal and occipital regions, whereas the control cases exhibited activations in the middle regions of the brain, including frontal, parietal, and temporal regions. This offered valuable insights into the discriminative features associated with different MS subgroups.

2.1.2. Methods for Classification of MS using Imaging Data and Clinical Data

These methods explore the integration of imaging and clinical data for the classification of MS groups.

The study of Vatian et al. (2019) for the classification between MS and no MS patients, combined clinical reports information with imaging data. The clinical report module employs Bidirectional Encoder Representations from Transformers (BERT) and Long Short-Term Memory networks to process textual information. For the image processing module, they use U-Net and VGG11 networks. In the final stage, they test early and late fusion of information where the outputs of BERT and the CNNs were concatenated and fed into a set of fully connected (FC) layers. The results showed a significant improvement in classification accuracy, with the method

achieving 80% accuracy compared to a baseline of 60% when only using images. This demonstrated the effectiveness of integrating textual and imaging data through deep learning models for diagnosing MS.

Another approach was proposed by Yoo et al. (2019), for predicting the conversion from CIS to MS, and it used neural networks to extract latent information of the MS lesions by computing Euclidean Distance Transform masks, which indicate the distance to the closest lesion boundary per pixel. They also fed 11 user-defined measurements composed of clinical general information, EDSS, and volumetric ratios of the brain. For the merging stage, they concatenated the information before the final FC layer.

Similarly, Roca et al. (2020) presented an algorithm that used MRI images (FLAIR, T1) and clinical data (sex, age, volume of lateral ventricles) to predict EDSS score. They employed a patch-based CNN and machine learning strategies to extract features from these inputs. Ultimately, the feature information from the clinical data was combined before the last FC layer of the network.

2.1.3. Methods for Combining Imaging Data and Clinical Data in other Disease Domain

The previously described methods are cases found in the literature for the classification of MS groups using imaging and clinical data. These methods used more naive strategies to combine the information by concatenating the features before passing to the FC layers of the network, or in the case of early fusion of Vatian et al. (2019), before the subnetwork of FC layers.

However, more exploration over combining these types of data has been applied in the classification or prediction of Alzheimer's disease or prognosis of cancer with histopathological images, clinical and genetic data. It has been shown that both information sources contribute to the diagnosis step.

The more common approach to information introduction from these two data forms is to concatenate feature vectors before passing through the last FC layer of the CNN architecture, as shown before in some studies related to the MS groups and to another commonly studied neurodegenerative disease, the Alzheimer's (Hao et al., 2019; Kopper et al., 2021; Pölsterl et al., 2020).

To achieve a more balanced combination of clinical and imaging data, El-Sappagh et al. (2020), Li et al. (2020b), Mobadersany et al. (2018) and Spasov et al. (2019), incorporated a multi-layer perceptron (MLP) instead of the last FC layer. This modification allowed for a non-linear contribution between the two modalities, facilitating a more balanced integration of clinical and imaging data within the model. The previous approach was put into practice with works related to histological and genomic data, as well as Alzheimer's disease classification. However, the approach misses the interaction at the local or pixel level of the image with the tabular information.

Another approach was proposed by Braman et al. (2021) using attention-gate tensor fusion, to fuse the latent representation of radiologic, pathologic, and genomic data, showing an improvement in the classification of glioma patients.

Duanmu et al. (2020) used multiplicative fusion by utilising an auxiliar network to generate a scalar scaling factor from the tabular information to rescale the feature maps. This generated a latent image representation dependent on the corresponding tabular data. However, this approach increases the runtime and memory requirements.

Finally, Wolf et al. (2022) proposed a Dynamic Affine Features Map Transform (DAFT), a general-purpose model that gets high-level concepts from the 3D images using feature maps of a convolutional layer on patients' images and tabular information.

Given the previous combination strategies of clinical data presented in this section, the best results are obtained by the DAFT approach (Wolf et al., 2022). The validation of this method includes classification and survival experiments conducted using Alzheimer's disease data from the ADNI[1] (Jack Jr et al., 2008). In both scenarios, the DAFT strategy consistently yielded superior results.

3. Material and methods

3.1. Dataset

In this study, 86 MS patients divided into four clinical profiles (12 CIS, 29 RRMS, 26 SPMS, 19 PPMS) were used. Each patient underwent several consecutive MR examinations, six on average, resulting in one scan per visit, using a 1.5T Siemens Sonata system (Siemens Medical Solution, Erlangen, Germany) with 8-channel head-coil. For a thorough description of the dataset, we refer to Kocevar et al. (2016).

We use the output of ico**brain ms** (FDA-approved an CE-cleared software) (Rakić et al., 2021), with T1, FLAIR images, lesion masks as well as tabular data of different time points including the EDSS score and MS group for each patient. These will be used as labels for classification tasks. The T1 and FLAIR images of the patients are preprocessed by registering them to the MNI space and bias field removal was performed on the T1 images using N4 (Tustison et al., 2010).

[1] The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

	CIS	RR	SP	PP
Subset 1	3	7	6	4
Subset 2	3	7	6	5
Subset 3	3	7	7	5
Subset 4	3	8	7	5

Table 1: Table representing the manual distribution of data for the 3-fold cross-validation in terms of MS groups.

Name	EDSS
low EDSS	0 – 3.5
medium EDSS	4 – 4.5
high EDSS	5 – 9.5

Table 2: Table representing the discretisation of EDSS scores for classification purposes.

3.2. Data Splitting

Regarding the dataset distribution for the experiments, we decided to use a 3-fold cross-validation. The folds will maintain the distribution of patients over the different types of experiments for comparison purposes. We manually divided the data into 4 subsets of patients (20, 21, 22, and 23 cases respectively). These subsets contain a stratified sample of patients, taking into account their MS groups and EDSS scores. In each fold, we used two subsets for the training set, one as an evaluation set and one that will be the same over the folds as test set (subset 4).

3.3. Classification of MS Groups and EDSS scores

We used the strategies that will be presented in the next section for the classification task. Furthermore, we will also present experiments concerning some of the previously mentioned strategies and approaches that involve just images or tabular information for comparability proposes.

The experiments entail three classification tasks associated with MS groups or EDSS score. These tasks encompass distinguishing patients between mild disease patterns and progressive disease evolution in two groups [CIS-RR, SP-PP] and classifying them into four groups [CIS, RR, SP, PP]. Also, we perform a discrete version of the EDSS score outlined in Table 2. The range of EDSS scores shown in the table is based on the distribution of cases present in the dataset so as to have 3 relevant and equilibrated ranges.

3.3.1. DAFT: Dynamic Affine Feature Map Transform

We use the DAFT open-source repository as a starting point for the classification task (Wolf et al., 2022). Initially, the network with the DAFT block receives a 3D T1-weighted image.

It was shown that the use of lesion labels as additional inputs of the network substantially improved the accuracy when predicting MS patients (Sepahvand et al., 2019). Therefore, we modify the architecture to receive 3 channels of 3D images. The new architecture is shown in Fig. 1. The changes imply:

1. Modification in the computation of normalisation options in the repository as well as the use of the binary mask of the brain for the standardisation option.
2. Implementation of the Hierarchical Data Format version 5 (HDF5) file (The HDF Group, 1997-NNNN) to use tabular, as well as imaging information in each epoch. The authors of DAFT use the HDF5 file that stores both of them in a per-patient manner using groups and datasets. The final structure can be seen in Fig. 2.
3. For experimentation proposes we add several schedulers and activation functions in the code.
4. To avoid overfitting, we use dropout and augmentation (flipping, injecting Gaussian noise).
5. Zhang et al. (2018) shows that using PReLU improves the model fitting with almost no change in the computational cost. We implemented this approach in our work.

Also, since the CIS and early RR scans share a lot of characteristics and are not easily distinguishable from one another, and moreover the same can be said about the relation of PP and SP scans, we performed the binary classification experiments. In these experiments, we treat the CIS and RR subjects as one class (roughly corresponding to mild disease patterns), whereas the PP and SP subjects form the second class (indicative of progressive disease evolution).

Additionally, we perform a classification task also with a discrete representation of the EDSS score, ranging the scores as shown in Table 2.

3.3.2. Siamese Networks

Originally used for signature verification systems (Bromley et al., 1993), Siamese networks receive a pair of images into equivalent encoders. Nonetheless, in the literature, they have been used for medical imaging proposes on different occasions (Birenbaum and Greenspan, 2016; Denner et al., 2021). In the Siamese architecture, it is proposed to differentiate between different groups by feeding each of the inputs with an image of each group. This has been applied in other neurodegenerative images, such as Alzheimer's disease, but not in multiple sclerosis images. Moreover, it has also been suggested to work when predicting the progress of a disease (Li et al., 2020a).

3.3.3. Modification over Siamese Network with DAFT block

Our objective is to demonstrate the significance of integrating diverse data types, which contain pertinent information and biomarkers, for the purpose of classify-

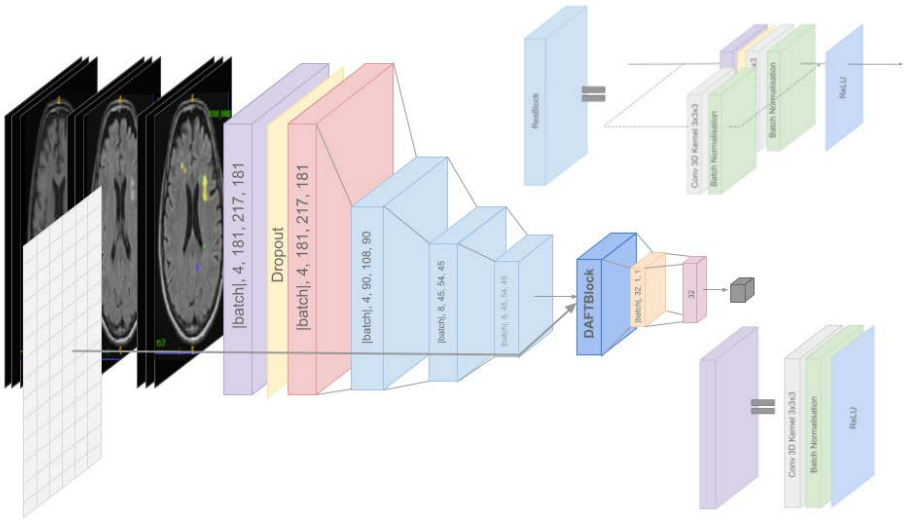


Figure 1: The figure illustrates the modified structure of DAFT with ResNet as its backbone. The architecture input is a concatenation of a binary version of the lesion mask, FLAIR and T1 images of the patient, along with tabular information. It generates a Tensor of variable dimensions, depending on the configured experiments.

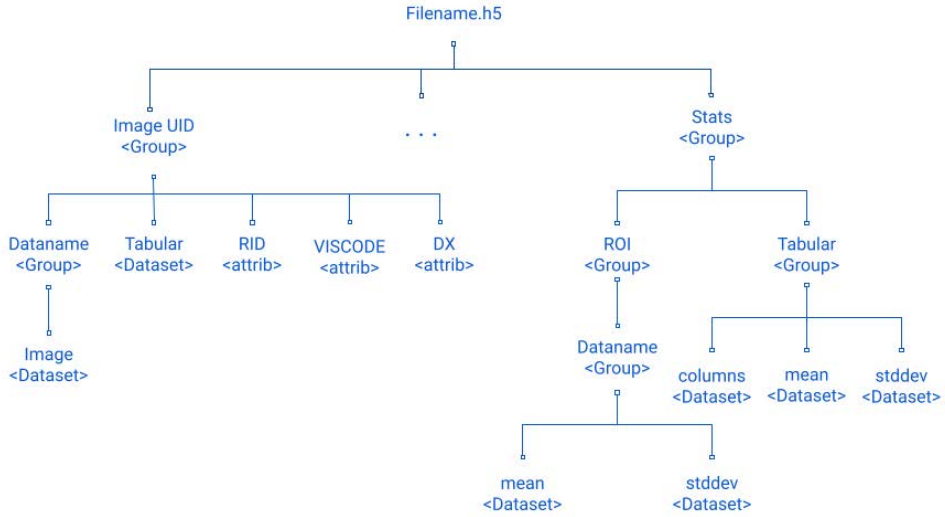


Figure 2: Representation of the HDF5 file structure for loading the dataset.

ing MS groups. Acknowledging the advantages offered by Siamese networks in this classification task, we have modified the encoder section of the Siamese architecture.

Specifically, we incorporate the encoder as the DAFT network we use for the classification task experiments. The previous approach was to incorporate the DAFT block into the encoder. In that way, we give the ability to mix information of imaging and tabular nature.

Additionally, we will employ the Euclidean distance metric with the contrastive loss to quantify the similarity between the two latent feature representations. These modifications aim to enhance the performance and accuracy of our model for MS group classification.

3.3.4. Implementation Details

We implemented all the previously described steps using Python 3.7.16 as well as relevant libraries such as PyTorch 1.12.1, and many others listed in the requirements file. The rest of the libraries and packages, as well as the code, are available in our GitHub repository (https://github.com/emyesme/DAFT).

3.3.5. Evaluation analysis and measures

Given the characteristics of the problem of classifying MS groups in an imbalanced dataset of patients, we opted to use balanced accuracy and confusion matrices to assess the results in the test set of each fold. Balanced accuracy is an appropriate metric for an im-

balanced dataset since performs the average accuracy for each class considering the majority and minority classes' accuracy. Also, we will compute the confusion matrices to examine the distribution of true positives, true negatives, false positives, and false negatives for each class, enabling a more detailed examination of the model's strengths and weaknesses in correctly classifying instances.

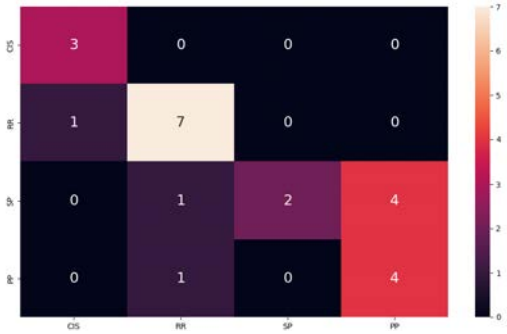
Due to the implementation of the 3-fold cross-validation strategy in our experiments, we will obtain a balanced accuracy and a confusion matrix for each fold. To generate the most optimal decision based on the three trained models, we implemented soft voting by combining the probabilities of prediction from each model from each fold and picking the prediction with the highest total probability. From now on, we will refer to these results as the ensemble balanced accuracy and ensemble confusion matrix, respectively.

4. Results

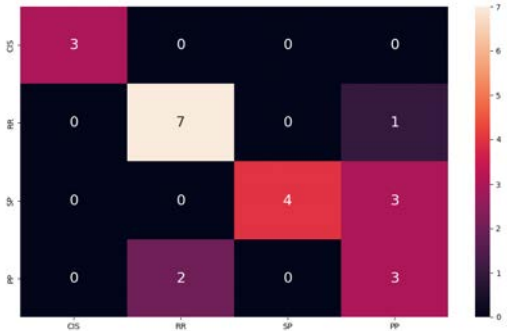
In this section, we present the outcomes of employing the DAFT baseline and variations of it. Initially, we evaluate DAFT to combine tabular and imaging data for MS classification in the authors' configuration. The results are presented in Table 3. Additionally, the table includes the initially modified version of DAFT for comparison purposes. In most cases, our modified DAFT exhibits higher mean balanced accuracy. However, we can notice more variability in the standard deviation and median. This situation may be explained by data imbalance and sensitivity to different input samples. The gap between training and validation performance that the variability produces can be addressed with regularisation strategies.

Subsequently, considering this challenge, we apply regularisation strategies, such as dropout and augmentation. Initially, when implementing the dropout strategy, we consider the related work in the state-of-the-art and conducted experiments varying the dropout probability within the range of [0.4, 0.7] with a step size of 0.1. The results are presented in Table 4. The most favorable results are at probabilities of 0.4 and 0.7. Considering the performance in the three experiments, the most favorable result is obtained using probability 0.6. From now on, all subsequent experiments that utilise dropout will employ this specific probability value. When incorporating both dropout and augmentation the results are shown in Table 5. The results demonstrate improvement in experiments with the implementation of these strategies.

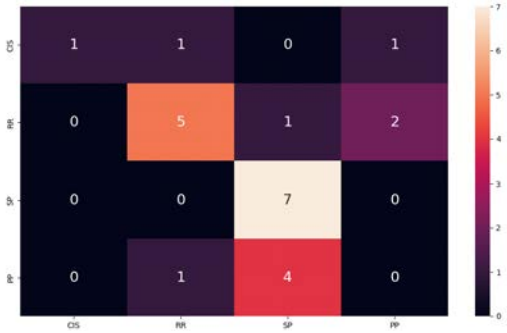
PReLU has demonstrated promising results in the binary classification task between MS and non-MS patients as shown in the state-of-the-art section. Therefore, we present an experiment varying the activation function from the original configuration of the author with ReLU to PReLU. The results shown in Table 6, for



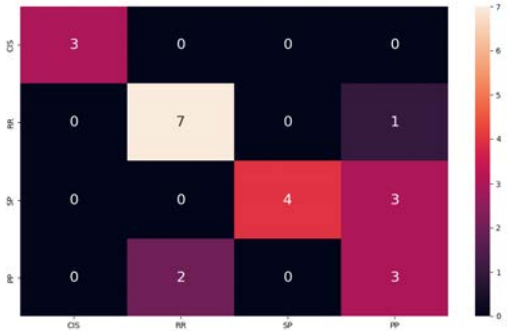
(a) No dropout, no augmentation



(b) No augmentation, but with dropout



(c) No dropout, but with augmentation



(d) With both dropout and augmentation

Figure 3: Confusion matrices based on the ensemble of the models obtained on the 3 folds for the 4 groups experiment. In each matrix, the rows correspond to the predicted classes and the columns to the ground truth classes.

Table 3: Performance results of unimodal and multimodal input DAFT approaches. This is computed over the test set. Mean and std refer to the mean and standard deviation of the results of each fold. 2 groups: [CIS-RR, PP-SP], 4 groups: [CIS, RR, PP, SP], 3 EDSS: [0-3.5, 4-4.5, 5+].

Experiment	**T1**		**T1 + FLAIR + lesion mask**	
	mean (std)	**Ensemble**	**mean (std)**	**Ensemble**
MS Subtypes Classification				
2 groups	0.83 (0.01)	**0.87**	0.77 (0.09)	0.79
4 groups	0.38 (0.06)	0.50	0.66 (0.07)	**0.63**
EDSS Classification				
3 EDSS	0.46 (0.13)	0.47	0.55 (0.08)	**0.56**

Table 4: Classification results using our modified DAFT approach varying the dropout probability in the range $[0.4, 0.7]$. This is computed over the test set. Mean and std refer to the mean and std of the results of each fold. 2 groups: [CIS-RR, PP-SP], 4 groups: [CIS, RR, PP, SP], 3 EDSS: [0-3.5, 4-4.5, 5+].

Experiments	**Dropout probability**							
	p=0.4		**p=0.5**		**p=0.6**		**p=0.7**	
	mean(std)	**Ensemble**	**mean(std)**	**Ensemble**	**mean(std)**	**Ensemble**	**mean(std)**	**Ensemble**
MS Subtypes Classification								
2 groups	0.76 (0.07)	**0.87**	0.80 (0.07)	0.83	0.79 (0.06)	**0.87**	0.82 (0.05)	0.79
4 groups	0.55 (0.04)	0.60	0.61 (0.07)	0.66	0.63 (0.12)	**0.76**	0.64 (0.08)	0.63
MS EDSS Classification								
3 edss	0.52 (0.09)	**0.60**	0.54 (0.07)	0.56	0.54 (0.07)	**0.60**	0.53 (0.05)	0.56

Table 5: Classification results using our modified DAFT approach implementing different balancing strategies. This is computed over the test set. Mean and std refer to the mean and std of the results of each fold. 2 groups: [CIS-RR, PP-SP], 4 groups: [CIS, RR, PP, SP], 3 EDSS: [0-3.5, 4-4.5, 5+].

Dropout	**Augmentation**	**Balanced Accuracy**	
		mean (std)	**Ensemble**
2 groups			
–	–	0.77 (0.09)	0.79
✓	–	0.79 (0.06)	0.87
–	✓	0.79 (0.02)	0.83
✓	✓	0.79 (0.06)	**0.87**
4 groups			
–	–	0.66 (0.07)	0.63
✓	–	0.63 (0.12)	0.76
–	✓	0.59 (0.10)	0.58
✓	✓	0.63 (0.13)	**0.76**
3 EDSS			
–	–	0.55 (0.08)	0.56
✓	–	0.54 (0.07)	0.60
–	✓	0.50(0.04)	0.53
✓	✓	0.54 (0.07)	**0.60**

Table 6: Classification results on our modified DAFT approach applying ReLU original configuration of the author against PReLU. This is computed over the test set. Mean and std refer to the mean and standard deviation of the results of each fold. "aug" refers to adding the augmentation strategie. 2 groups: [CIS-RR, PP-SP], 4 groups: [CIS, RR, PP, SP], 3 EDSS: [0-3.5, 4-4.5, 5+].

Experiment	**p = 0.6 + aug**		**p = 0.6 + aug + PReLU**	
	mean (std)	**Ensemble**	**mean (std)**	**Ensemble**
MS Subtypes Classification				
2 groups	0.79 (0.06)	0.87	0.80 (0.07)	**0.92**
4 groups	0.63 (0.13)	**0.76**	0.65 (0.07)	0.62
MS EDSS Classification				
3 EDSS	0.54 (0.07)	**0.60**	0.52 (0.07)	0.53

Table 7: Final classification results of experiments over the test set. Mean and std refer to mean and std of the results of each fold. 2 groups: [CIS-RR, PP-SP], 4 groups: [CIS, RR, PP, SP], 3 EDSS: [0-3.5, 4-4.5, 5+].

	no DAFT		**RF**		**DAFT**	
Experiments	**mean (std)**	**Ensemble**	**mean (std)**	**Ensemble**	**mean (std)**	**Ensemble**
MS Subtypes Classification						
2 Groups	0.67(0.15)	0.70	0.94 (0.02)	**0.92**	0.80 (0.07)	**0.92**
4 Groups	0.37 (0.02)	0.35	0.43 (0.10)	0.58	0.63 (0.13)	**0.76**
MS EDSS Classification						
3 EDSS	0.41 (0.10)	0.46	0.54(0.01)	0.56	0.54 (0.07)	**0.60**

the classification of the two groups exhibit high ensemble balanced accuracy. Nonetheless, for the remaining two experiments, the variation does not outperform the current outcomes obtained using the ReLU activation function.

In Table 7, we provide a final comparison between the best-performing modified DAFT configuration, a configuration employing the same architecture but excluding the DAFT block (receiving only imaging data as input), and a well-known machine learning strategy, namely random forest, which utilises the tabular information. In general, the modified DAFT strategy has the most favorable results. However, in the two-group experiment, the random forest method outperforms the modified DAFT strategy. This observation aligns with the understanding that machine learning strategies can outperform deep learning approaches depending on the dataset size (Shwartz-Ziv and Armon, 2022).

Regarding the utilisation of the Siamese approach with the DAFT block, we specifically implemented this approach to distinguish between mild disease patterns and progressive disease evolution in two groups, namely [CIS-RR, SP-PP], due to the inherent nature of the architecture. In this particular case, we achieved a mean accuracy and standard deviation of 0.61 and 0.06 respectively, along with an ensemble accuracy of 0.63. However, it is worth noting that the accuracy curve exhibited volatility during both the training and evaluation stages. As a result, we did not conduct any further experiments using this architecture.

The observed fluctuation in accuracy suggests that the architecture exhibits uncertainty when making decisive choices, depending on the specific fold it is trained on. This behavior is to be expected, given the sensitivity of the architecture to certain input samples, as demonstrated in our other experiments.

5. Discussion

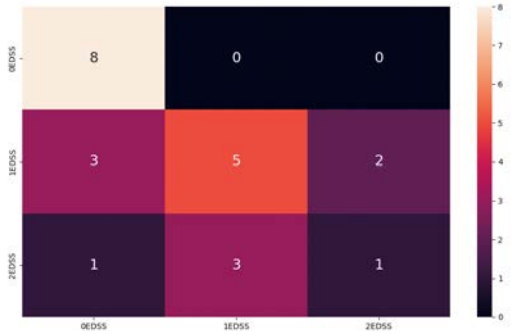
In this study, our objective is to investigate the potential of a deep learning strategy, namely DAFT, which combines medical imaging and clinical information to classify MS groups and EDSS scores. Previous research has demonstrated the effectiveness of this strategy in improving the classification and disease progression prediction for Alzheimer's disease.

To evaluate the performance of the DAFT architecture in the context of MS classification, we conducted a baseline experiment comparing it with our modified approach that incorporates state-of-the-art techniques. Specifically, we concatenated imaging sequences (T1, FLAIR) and a lesion mask to enhance the classification of MS groups. In Table 3, we observe that our modification generally improves the results compared to the previous DAFT architecture. However, the variability of our results is higher due to the inherent imbalance in the problem. To address this issue, we implemented several strategies inspired by the literature (Wang et al., 2018; Zhang et al., 2018).

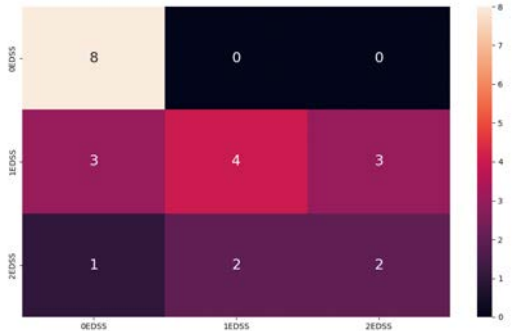
Table 4 showcases the results of varying the dropout probability within the range of [0.4, 0.7]. Given that in three experiments there is a higher mean and ensemble balanced accuracy from the dropout probability of 0.6 we select the performance. Subsequently, in Table 5, we conducted experiments to investigate the impact of the augmentation strategy. The combination of dropout and augmentation, as well as the implementation of only the dropout strategy reach the best results in all the metrics. In certain instances, the addition of augmentation without dropout results in a decrease in the overall performance of balanced accuracy. This phenomenon can occur when there is a high sensitivity to specific input samples.

By interpreting the confusion matrices presented in Fig. 3, overall, we can notice the network has problems deciding over the progressive MS groups [SP, PP] while the mild disease groups [CIS, RR] are classified with more certainty. Specifically, the confusion matrix corresponding to the best-balanced accuracy, item (d), presents more problems to distinguish between progressive disease groups [SP, PP]. This can be related to the fact that these groups share several MRI features such as lesion distribution, brain atrophy patterns, etc.

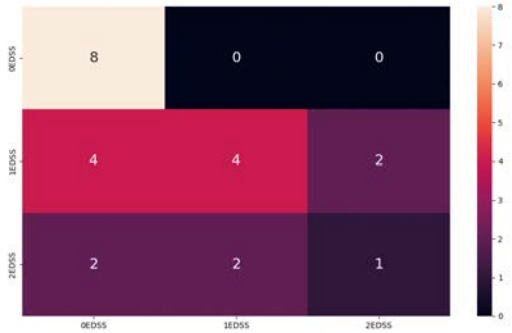
Let us focus on the best confusion matrix that corresponds to the highest balanced accuracy for the experiment 4 groups, shown in Fig. 3 item (d). The matrix showcases difficulty in distinguishing between SP and PP. This can be attributed to the fact that these groups share numerous MRI features, which may lead to over-



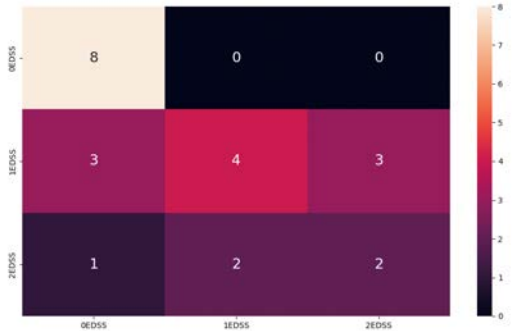
(a) No dropout, no augmentation



(b) No augmentation, but with dropout

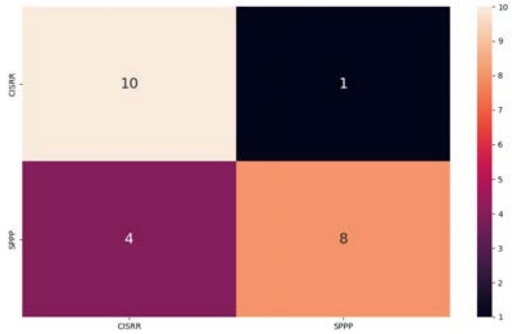


(c) No dropout, but with augmentation

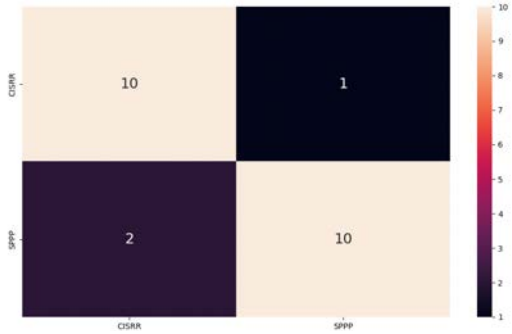


(d) With both dropout and augmentation

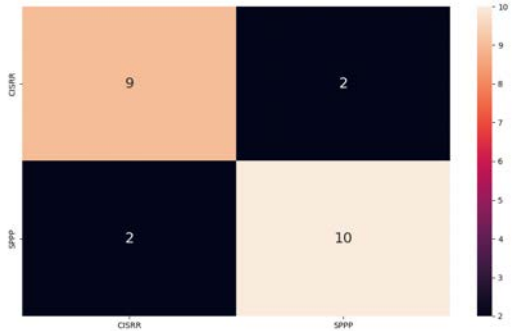
Figure 4: Confusion matrices based on the ensemble of the models obtained on the 3 folds for the EDSS discretised experiment. In each matrix, the rows correspond to the predicted classes and the columns to the ground truth classes.



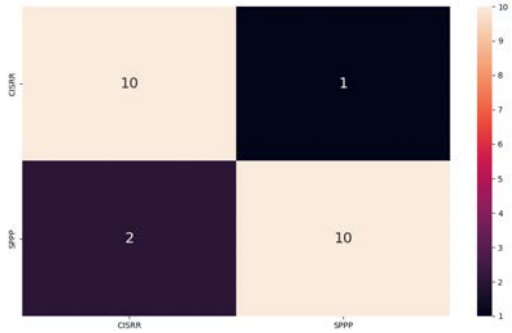
(a) No dropout, no augmentation



(b) No augmentation, but with dropout



(c) No dropout, but with augmentation



(d) With both dropout and augmentation

Figure 5: Confusion matrices based on the ensemble of the models obtained on the 3 folds for the 2 group experiment. In each matrix, the rows correspond to the predicted classes and the columns to the ground truth classes.

lapping patterns in the extracted information.

Due to these shared features, the network may struggle to capture the subtle differences between the two progressive disease groups, resulting in higher misclassification rates and reduced certainty in the classification outcomes.

In the experiment concerning the discretised EDSS, the findings from the confusion matrices depicted in Fig. 5 show the network's ability to effectively discriminate between different EDSS categories. The network demonstrates a remarkable proficiency in distinguishing low EDSS scores from those categorised as middle or high. This observation suggests the presence of distinct MRI features that serve as reliable indicators for differentiating the more disabling EDSS scores from the rest.

The network's capacity to accurately discern between low EDSS scores and higher disability levels attests to the robustness and discriminative capabilities of the model. This distinction enables the identification and characterisation of patients with more severe disability. The fact that the network can readily differentiate these debilitating EDSS scores from others signifies the existence of pronounced MRI features associated with higher disability levels.

Regarding the Siamese approach, the conducted experiment for classification yielded relatively low accuracy. Nevertheless, existing literature (Birenbaum and Greenspan, 2016; Denner et al., 2021) suggests that this type of architecture is valuable for predicting disability scores. Therefore, exploring its potential for the prediction task could be a promising avenue for future research.

5.0.1. Limitations

Our work has limitations, and it is crucial to acknowledge these constraints in order to provide a comprehensive understanding of the study:

Firstly, one notable limitation is the absence of detailed records regarding treatment administration or its absence. This lack of information prevents us from studying the potential confounding factors related to therapeutic effects.

Secondly, there are recognised biomarkers for MS in blood test results or genetic data. Incorporating these additional factors could provide valuable insights and potentially enhance the predictive capabilities of our model.

By addressing these suggestions, we can further enhance the accuracy and reliability of MS classification and prognosis models, ultimately improving patient care and treatment decision-making

6. Conclusions

In summary, the use of deep learning strategies, specifically DAFT, to combine multiple medical imaging sequences with lesion masks, and clinical information leads to improved classification of MS groups and EDSS scores compared to the individual use of medical imaging or clinical information.

It is worth noting that the choice of utilising a single form of information depends on factors such as sample size, the relevance of the available clinical data, and the quality of medical imaging acquisition protocols. In certain cases, employing a method that exclusively leverages one type of information can be a prudent decision.

This study focuses on the classification of MS groups and EDSS scores, yielding favorable results. However, the same methods can be applied to predict the EDSS score of a patient over a given time period. In the case of the Siamese architecture, instead of inputting images and clinical information from patients across different MS groups to each encoder, the input would consist of at least two time points of the same patient, incorporating the EDSS score from time point 1 and predicting for the same patient at time point 2.

Acknowledgments

Thank you MAIA, thank you icometrix, thank you supervisors, thank you family and friends.

References

Aslam, N., Khan, I.U., Bashamakh, A., Alghool, F.A., Aboulnour, M., Alsuwayan, N.M., Alturaif, R.K., Brahimi, S., Aljameel, S.S., Al Ghamdi, K., 2022. Multiple sclerosis diagnosis using machine learning and deep learning: Challenges and opportunities. Sensors 22, 7856.

Birenbaum, A., Greenspan, H., 2016. Longitudinal multiple sclerosis lesion segmentation using multi-view convolutional neural networks, in: Deep Learning and Data Labeling for Medical Applications: First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings 1, Springer. pp. 58–67.

Braman, N., Gordon, J.W., Goossens, E.T., Willis, C., Stumpe, M.C., Venkataraman, J., 2021. Deep orthogonal fusion: multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24, Springer. pp. 667–677.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R., 1993. Signature verification using a "siamese" time delay neural network. Advances in neural information processing systems 6.

Calimeri, F., Marzullo, A., Stamile, C., Terracina, G., 2018. Graph based neural networks for automatic classification of multiple sclerosis clinical courses., in: ESANN.

Denner, S., Khakzar, A., Sajid, M., Saleh, M., Spiclin, Z., Kim, S.T., Navab, N., 2021. Spatio-temporal learning from longitudinal data for multiple sclerosis lesion segmentation, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part I 6, Springer. pp. 111–121.

Duanmu, H., Huang, P.B., Brahmavar, S., Lin, S., Ren, T., Kong, J., Wang, F., Duong, T.Q., 2020. Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using deep learning with integrative imaging, molecular and demographic data, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23, Springer. pp. 242–252.

Eijlers, A.J., Meijer, K.A., van Geest, Q., Geurts, J.J., Schoonheim, M.M., 2018. Determinants of cognitive impairment in patients with multiple sclerosis with and without atrophy. Radiology 288, 544–551.

Eitel, F., Soehler, E., Bellmann-Strobl, J., Brandt, A.U., Ruprecht, K., Giess, R.M., Kuchling, J., Asseyer, S., Weygandt, M., Haynes, J.D., et al., 2019. Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation. NeuroImage: Clinical 24, 102003.

El-Sappagh, S., Abuhmed, T., Islam, S.R., Kwak, K.S., 2020. Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data. Neurocomputing 412, 197–215.

Hao, J., Kosaraju, S.C., Tsaku, N.Z., Song, D.H., Kang, M., 2019. PAGE-Net: interpretable and integrative deep learning for survival analysis using histopathological images and genomic data, in: Pacific Symposium on Biocomputing 2020, World Scientific. pp. 355–366.

Jack Jr, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L. Whitwell, J., Ward, C., et al., 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine 27, 685–691.

Kocevar, G., Stamile, C., Hannoun, S., Cotton, F., Vukusic, S., Durand-Dubief, F., Sappey-Marinier, D., 2016. Graph theory-based brain connectivity for automatic classification of multiple sclerosis clinical courses. Frontiers in neuroscience 10, 478.

Kopper, P., Pölsterl, S., Wachinger, C., Bischl, B., Bender, A., Rügamer, D., 2021. Semi-structured deep piecewise exponential models, in: Survival Prediction-Algorithms, Challenges and Applications, PMLR. pp. 40–53.

Li, M.D., Chang, K., Bearce, B., Chang, C.Y., Huang, A.J., Campbell, J.P., Brown, J.M., Singh, P., Hoebel, K.V., Erdoğmuş, D., et al., 2020a. Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging. NPJ digital medicine 3, 48.

Li, S., Shi, H., Sui, D., Hao, A., Qin, H., 2020b. A novel pathological images and genomic data fusion framework for breast cancer survival prediction, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE. pp. 1384–1387.

Lublin, F.D., Reingold, S.C., Cohen, J.A., Cutter, G.R., Sørensen, P.S., Thompson, A.J., Wolinsky, J.S., Balcer, L.J., Banwell, B., Barkhof, F., et al., 2014. Defining the clinical course of multiple sclerosis: the 2013 revisions. Neurology 83, 278–286.

Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D.A., Barnholtz-Sloan, J.S., Velázquez Vega, J.E., Brat, D.J., Cooper, L.A., 2018. Predicting cancer outcomes from histology and genomics using convolutional networks. Proceedings of the National Academy of Sciences 115, E2970–E2979.

Pölsterl, S., Sarasua, I., Gutiérrez-Becker, B., Wachinger, C., 2020. A wide and deep neural network for survival analysis from anatomical shape and tabular clinical data, in: Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I, Springer. pp. 453–464.

Rakić, M., Vercruyssen, S., Van Eyndhoven, S., de la Rosa, E., Jain, S., Van Huffel, S., Maes, F., Smeets, D., Sima, D.M., 2021. icobrain ms 5.1: Combining unsupervised and supervised approaches for improving the detection of multiple sclerosis lesions. NeuroImage: Clinical 31, 102707.

Roca, P., Attye, A., Colas, L., Tucholka, A., Rubini, P., Cackowski, S., Ding, J., Budzik, J.F., Renard, F., Doyle, S., et al., 2020. Artificial intelligence to predict clinical disability in patients with multiple sclerosis using FLAIR MRI. Diagnostic and Interventional Imaging 101, 795–802.

Sepahvand, N.M., Hassner, T., Arnold, D.L., Arbel, T., 2019. CNN prediction of future disease activity for multiple sclerosis patients from baseline MRI and lesion labels, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4, Springer. pp. 57–69.

Shoeibi, A., Khodatars, M., Jafari, M., Moridian, P., Rezaei, M., Alizadehsani, R., Khozeimeh, F., Gorriz, J.M., Heras, J., Panahiazar, M., et al., 2021. Applications of deep learning techniques for automated multiple sclerosis detection using magnetic resonance imaging: A review. Computers in Biology and Medicine 136, 104697.

Shwartz-Ziv, R., Armon, A., 2022. Tabular data: Deep learning is not all you need. Information Fusion 81, 84–90.

Spasov, S., Passamonti, L., Duggento, A., Lio, P., Toschi, N., Initiative, A.D.N., et al., 2019. A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. Neuroimage 189, 276–287.

The HDF Group, 1997-NNNN. Hierarchical Data Format, version 5. Https://www.hdfgroup.org/HDF5/.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: improved N3 bias correction. IEEE transactions on medical imaging 29, 1310–1320.

Vatian, A., Gusarova, N., Dobrenko, N., Klochkov, A., Nigmatullin, N., Lobantsev, A., Shalyto, A., 2019. Fusing of medical images and reports in diagnostics of brain diseases, in: Proceedings of the 2019 the International Conference on Pattern Recognition and Artificial Intelligence, pp. 102–108.

Walton, C., King, R., Rechtman, L., Kaye, W., Leray, E., Marrie, R.A., Robertson, N., La Rocca, N., Uitdehaag, B., van Der Mei, I., et al., 2020. Rising prevalence of multiple sclerosis worldwide: Insights from the atlas of MS. Multiple Sclerosis Journal 26, 1816–1821.

Wang, S.H., Tang, C., Sun, J., Yang, J., Huang, C., Phillips, P., Zhang, Y.D., 2018. Multiple sclerosis identification by 14-layer convolutional neural network with batch normalization, dropout, and stochastic pooling. Frontiers in neuroscience 12, 818.

Wolf, T.N., Pölsterl, S., Wachinger, C., Initiative, A.D.N., et al., 2022. DAFT: a universal module to interweave tabular data and 3D images in CNNs. NeuroImage 260, 119505.

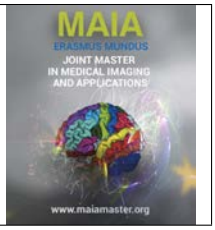
Yoo, Y., Tang, L.Y., Li, D.K., Metz, L., Kolind, S., Traboulsee, A.L., Tam, R.C., 2019. Deep learning of brain lesion patterns and user-defined clinical and MRI features for predicting conversion to multiple sclerosis from clinically isolated syndrome. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 7, 250–259.

Zhang, Y., Hong, D., McClement, D., Oladosu, O., Pridham, G., Slaney, G., 2021. Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. Journal of Neuroscience Methods 353, 109098.

Zhang, Y.D., Pan, C., Sun, J., Tang, C., 2018. Multiple sclerosis identification by convolutional neural network with dropout and parametric ReLU. Journal of computational science 28, 1–10.

Medical Imaging and Applications

Master Thesis, June 2023



Automatic Segmentation of histological images of the brain of mouse

Juan Cisneros[a], Alain Lalande (PhD)[a], Fabrice Meriaudeau (PhD)[a], Stephan Collins (PhD)[b]

[a]*ICMUB laboratory, University of Burgundy, Dijon, France*
[b]*NeuroGeMM, niversity of Burgundy, Dijon, France*

Abstract

The study of the mouse brain is of utmost importance in the field of neuroscience, as it broadly offers the best animal model for the study of the human brain. Specifically, genetical manipulation, so easily achieved in the mouse, allows us to explore the effects of genes on brain morphogenesis. The host laboratory has recently published a list of 198 genes through a high-throughput preclinical studies using using high-resolution and annually annotated histological images from thousands of mouse brains. Manual segmentation takes approximately 1 hour to 24 anatomical regions. This work consisted in producing an automatic system for the segmentation of these anatomical regions using the existing "ground truth" dataset. Deep learning methods were used based on a U-Net and a Attention U-Net architectures. This system was trained with about 2,000 annotated images for each region of interest. The average size of each image was 1 GB, thus one of the biggest challenges was to manage the volume of information in the images. Neuroanatomical regions differ in predictability such as the ventromedial nucleus of the hypothalamus (VHMvl) and the inferior colliculus (InfC). However, results show a 80.39% and 94.42% Dice scores respectively, making the deep learning extremely powerful for the annotation task. For the end-user, analyzing an image now consists of a 5 minute task, mainly through validation of automatically generated regions of interest

Keywords: Mouse brain, Anatomical phenotype, High resolution images, Segmentation, Deep Learning

1. Introduction

Neuroscience and the study of anatomical phenotypes are intricately linked areas of scientific research. Neuroscience is a multidisciplinary field that focuses on understanding the structure, function, and complex interactions within the nervous system, particularly the brain. Anatomical phenotypes, on the other hand, refer to the observable and measurable structural traits of an organism, which are influenced by both its genetic makeup and environmental factors. These phenotyes often refer to specific structural characteristics of the brain and other parts of the nervous system. This include macroscopic features, such as the size and shape of the different brain regions, and microscopic features, such as the organization and connectivity of neurons.

1.1. Neuroanatomical phenotype

The study of these neuroanatomical phenotypes provides valuable insights into how variations in brain structure relate to differences in function, behavior, and susceptibility to neurological disorders. It also allows researchers to investigate how specific genes and environmental factors influence the brain's physical characteristics. Qualitative assessments have been used (e.g., cerebellar agenesis, failure of the med-crossing of the corpus callosum), but the neuroanatomical phenotype varies in effect size.

Autism spectrum disorders (ASD) harbor a wide range of neuroanatomical phenotypes often considered as subtle and difficult to assess. Studying the neuroanatomy of autism thus requires defined phenotypes in different samples, at different stages of brain development with high resolution. This is why a very precise characterization of the regions of interest in the brain is necessary.

3D imaging techniques such as fMRI are complementary. Whilst the resolution achieved is less than standard histological techniques, these studies focus on examining patterns of functional connectivity, identifying specific brain regions or networks that show differences in activity or connectivity between individuals

with ASD and those without. These studies aim to uncover potential biomarkers, neural correlates and underlying mechanisms of ASD. (Hull et al., 2017).

The investigation of developmental diseases in human brains is complicated by reasons such as: obtaining patient consent, limited access to developing brains, procedures which must be non invasive, among others. Animal models are therefore essential for the study of developmental diseases. Working with animal models allows controlled experiments, where the environment, the age and the genetic background are controlled. Within a defined set of ethical guidelines, it is also possible to manipulate genes and carry out invasive procedures. Animal models also allow for longitudinal studies, have expanded sample sizes, and allow for a better understanding of the underlying mechanisms of developmental diseases (Bossert and Hagendorff, 2021).

1.2. Mouse Model

Whilst human studies of cognitive disorders and in particular, their genetic causes have had immense success in the last two decades, it is estimated that 50 percent of genetic cases remain unsolved. Often, because of partial penetrance status, unclear pattern of familiar inheritance, or mutations of unknowns significance. The mouse is now seen as a unique tool to validate genetic hypotheses and address the problem of missing heritability in genetic developmental disorders. Indeed, the human and mouse genomes share more than 90% of the sequence homology with almost the same number of genes (more than 20,000 genes) (Breschi et al., 2017).

Despite the obvious difference in size and some variations in brain structure, many fundamental aspects of brain organization, neuronal function and even behavior are conserved between mice and humans. But above all, for genetics studies, the mouse has the main advantage of being genetically homogeneous (when researchers use isogenic strains), offering a unique way to test whether a gain or loss of function at a specific locus in the genome is responsible for the same neuroanatomical phenotypes as a patient harboring a mutation in the same gene thereby providing the golden proof of causality.

Mice models then become invaluable in allowing extensive investigation into the function of a mutation and explain the pathophysiology of a disease. Changes in neuroanatomical phenotypes, such as atrophy of a particular brain region, can be indicative of disease progression in both mice and humans. Knowledge of these changes in mice allows researchers to better understand similar processes in humans, which in turn improve diagnostic tools, help anticipating disease progression and saves time in finding potential treatments. In Figure 1 is shown as an example key regions where fear memory is involved in humans and mice.

According to Breschi et al. (Breschi et al., 2017), the use of mouse models is very useful for the following reasons: a) The mouse life cycle is long enough to follow the evolution of a disease b) A lot of histological, anatomical or quantitative studies can be performed on the mouse brain, thanks to the fact that it can be extracted and applied to these various studies c) The mouse brain anatomy and physiology has great similarities with the human brain.

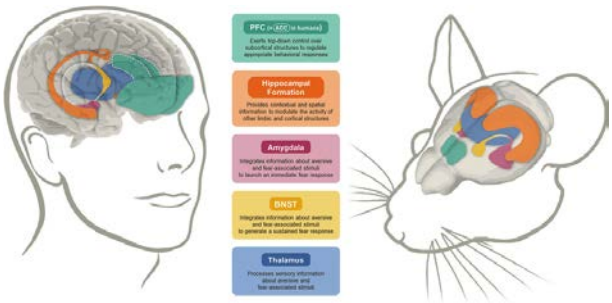


Figure 1: Key regions in the human and mouse brains involved in fear memory. (Flores et al., 2018)

Beyond this, the mouse brain's ability to be genetically manipulated allows for the exploration of the effects of genes on brain activity. These tiny creatures also serve as invaluable models for the study of development diseases such as ADHD (Majdak et al., 2016) and Autism spectrum disorder (Kazdoba et al., 2016), assisting in the discovery of potential treatments and the comprehension of the diseases' progression. With their practical size and rapid reproduction, mice are the ideal candidates for scientific investigation. Furthermore, the uniformity of the mouse brain reduces experimental error, additionaly enhancing the accuracy of findings. In essence, the examination of mouse brains is essential to the understanding of both healthy and diseased brain function, paving the way for potential therapies for neurological disorders.

The mouse offers a number of powerful tools to make an association between phenotypes and genotypes. An expanding repertoire of technologies exist to manipulate the mouse genome – to mutate, overexpress or knockout (KO) genes of interest, to help researchers study pathogenicity at a molecular, cellular, physiological and behavioral level. An example of the impact of these mutations is shown in Figure 2.

The NeuroGeMM laboratory, of the university of Burgundy, has identified 198 genes affecting brain morphogenesis through a high throughput screen of 1500 knock out lines (mouse lines where a specific gene has been inactivated) (Collins et al., 2019). The evaluation of anatomical abnormalities is typically done on high resolution histological images (24,000 pixels x 14,000 pixels sizes with 0.45 micrometer/pixel resolution). This task is carried out using manual annotation of the different regions of the mouse brain with semi-automatic software assisted methods.

The aim of this study is to produce an automatic sys-

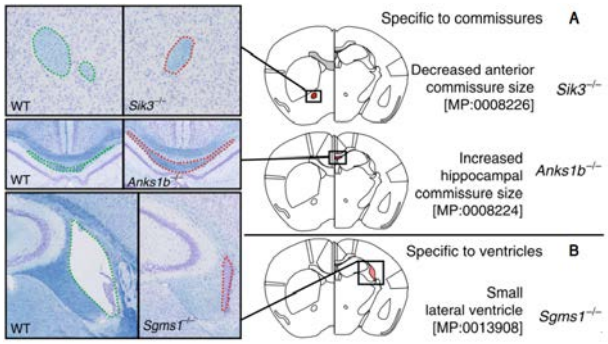


Figure 2: Impact of mutations on the mouse brain. Examples having a specific impact on: **A** the commissures and **B** the ventricles. (Collins et al., 2019)

tem for segmenting regions of interest in the mouse brain from histological images, with the purpose of avoiding manual methods and reducing the time dedicated to this task. Routinely, the task of annotating brain regions is not performed by expert personnel, because it is a repetitive and tedious task. It is assigned to students or laboratory assistants, which can result in a less reliable annotation and thus requires a significant amount of final checks for error detection.

In this paper we will evaluate the performance of different deep learning architectures for the task of segmenting regions of interest in the mouse brain. Some of the issues identified at the very beginning were the sheer size of images and images artefacts:

First, when working with histological images, image quality is of utmost importance. By having a greater detail of what is present in the image, abnormalities that may be imperceptible with a low resolution can be identified using careful analysis of cellular patterns. This is why histological slices (which have a 5 micrometer depth) were scanned at high resolution. For obvious reasons, dealing with such large images (1GB) is a main issue since reducing image quality meant risking lower segmentation quality.

Second, it is worth mentioning that not all regions are systematically taken by expert anatomist for any given image. Image quality (coloration, histological artefacts such as tears, folds or autolysis) and histological accuracies relative to precise stereotaxic coordinates are not always optimal. Hence, an expert may draw the hypothalamus for example, but not subcortical areas.

To overcome the above problems, we propose a model able to segment 24 mouse brain regions and its practical implementation in the laboratory. The main objective of the proposed method is the reduction of the time spent in segmenting the regions of the mouse brain. In the NeuroGeMM laboratory this activity takes about 1 hour for the manual annotation of all regions of a brain slice. In a normal working day, only 8 slices are fully annotated.

The work will be done mainly with 2D histological samples, as the laboratory collected data for more than 5 years. The present method proposes a general approach for the correct annotation of the 24 regions, which will serve as a starting point for future algorithms to specialize in different regions of interest. In addition, it will serve for a possible analysis of brain regions in new 3D samples that are currently being taken and manually annotated.

2. State of the art

The state of the art in medical image segmentation is driven by ongoing research and innovation, aiming to improve accuracy, efficiency, and generalizability across various medical imaging modalities and applications. With these advances, physicians have the ability to improve decision making, enhance patient care, and accelerate medical research and diagnosis. In this section, segmentation models will be explained, starting with the manual ones and ending with the fully automatic ones. Mainly for the segmentation of histological images of mouse brain.

2.1. Manual segmentation

Manual segmentation of medical images involves a human expert manually delineating structures or regions of interest on the images. Experience and precision are required to accurately delineate the boundaries of anatomical structures or lesions using specialized tools. This approach is an absolute prerequisite for evaluating automated or semi-automated methods. It is considered as the gold standard in cases requiring high precision and accurate delineation, such as radiation therapy, disease diagnosis or image-guided interventions.

The software mainly used to perform this activity are:

- ITK-SNAP is a program application used for segmentation, visualization and study of images in the field of medical imaging. With ITK-SNAP, users have the possibility of doing work such as manual or semi-automatic segmentation, volume representation, 3D visualization and quantitative study of medical images (Yushkevich et al., 2006).

- 3D-slicer is an open-source software platform for medical image analysis and visualization. It supports various types of medical imaging data and provides a wide range of tools and modules for tasks such as segmentation, registration, and volume rendering. 3D Slicer is widely used in research and education to enhance the exploration and analysis of medical images for improved diagnosis, treatment planning, and scientific investigation (Fedorov et al., 2012).

- ImageJ/Fiji is a popular open-source software package for image analysis and processing in the

life sciences. It offers a user-friendly interface and a broad range of tools for tasks such as enhancing, segmenting, quantifying, and visualizing images. With support for various image formats and a vast collection of plugins and macros, Fiji/ImageJ is widely used in research labs and academic institutions for biological and medical image analysis tasks (Schindelin et al., 2012).

- OsiriX is an advanced open-source software program designed specifically for navigating and viewing medical images. Developed by radiologists, it provides a solution for managing, interpreting and sharing radiological images. The software supports a wide range of medical imaging formats, including DICOM (Digital Imaging and Communications in Medicine), and enables 2D, 3D and 4D imaging, assisting healthcare professionals in diagnosis and treatment planning (Rosset et al., 2004).

For working with histopathological images, the most used is Fiji, because of the characteristics mentioned above, and for being specialized in 2D image processing.

2.2. *Semi-automatic segmentation*

Semi-automated segmentation combines manual interaction with automated algorithms. The advantage of this model is that the user provides the initial information and the algorithm refines the segmentation iteratively. This approach takes advantage of the user's experience while benefiting from automation, with the goal of obtaining accurate and efficient results with reduced manual effort. The process involves refining the segmentation based on user feedback, allowing for adjustments and validations as needed. Semi-automated segmentation strikes a balance between user input and automation to improve segmentation accuracy, efficiency and consistency. This method of segmentation can be performed by applying specialized image processing software, such as those mentioned in the previous section. Initialization can be done by dropping seed points, drawing manually or with some tools like thresholding and edge detection to start segmentation. Additionally there are segmentation refinement tools like: "Wand" or "Brush" for Fiji, "Region Growing" or "Live Wire" in ITK-SNAP, or "Paint" and "Grown for seeds" from 3D sclier. However, this automatic segmentation can also be carried out using deep learning methods. This approach maintain the main idea of reducing human intervention while increasing efficiency and accuracy in the recognition and correct segmentation of regions of interest.

Di Scandalea et al. (Di Scandalea et al., 2019) have developed Deep Active Learning, an open source Python-based simulation framework designed to segment myelin from histological data using uncertainty sampling. It uses the Keras framework and is based on a convolutional neural network architecture. The framework classifies pixels as myelin or background, providing a valuable tool for histological image analysis. The pipeline of the framework is presented in Figure 3. This active learning combines human experience with deep learning to iteratively select the most informative samples for annotation. Initially, a small set of labeled data is used to train a deep learning model. Next, the model is used to predict segmentations on unlabeled data, and the most uncertain or difficult samples are selected for manual annotation. This process continues iteratively, gradually improving model performance with the incorporation of additional labeled samples.

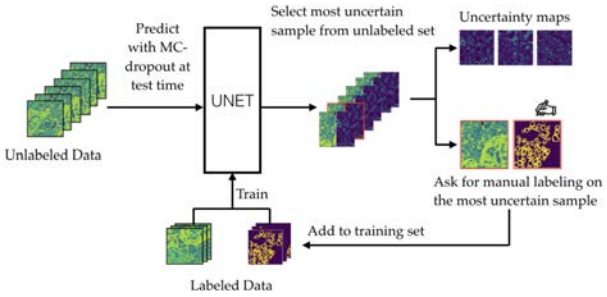


Figure 3: Deep Active Learning pipeline for semi-automatic segmentation histological images with manual annotation for uncertain sample to further data labeling (Di Scandalea et al., 2019)

The proposed interactive segmentation network from Jahanifar et al. (Jahanifar et al., 2021), offers an efficient approach for annotating various tissue types in histological images with minimal user input, framework shown in Figure 4. Users simply need to draw a few pixels within each region of interest as a guide signal for the model. To handle the diverse appearance and irregular geometry of different tissue regions, the network incorporates automatic and minimalistic techniques for generating guide signals. These techniques enhance the model's robustness to variations in user input, resulting in accurate and reliable segmentation. They use an EfficientNet network with and an extra Residual Multi-scale (RMS) block. The dataset used for this work contains 151 regions of H&E-stained tissue images extracted from WSI of as many triple-negative breast cancer cases acquired from the Cancer Genome Atlas.

These examples demonstrate how deep learning can be combined with user input to achieve semi-automatic segmentation, providing efficient and accurate segmentation results while allowing for user interaction and control in the segmentation process. Current advances have helped to develop new approaches to this type of semi-automatic segmentation, such as zero-shot segmentation.

Zero-shot segmentation is a computer vision technique propose by Bucher et al. (Bucher et al., 2019), where objects or regions in an image are segmented

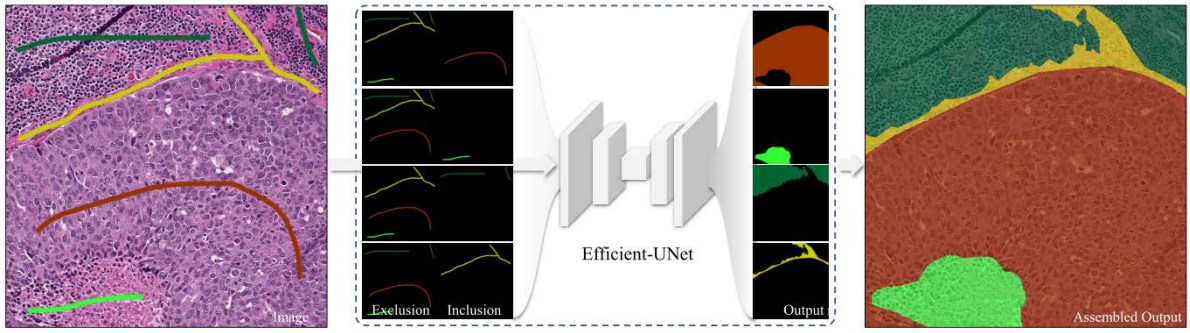


Figure 4: Proposed framework pipeline in Jahanifar et al. for histological images segmentation with minimal user input. (Jahanifar et al., 2021)

without the need for pre-training on specific classes. Instead, it relies on additional information, such as textual descriptions or attributes, to generalize the segmentation to unseen classes.

The idea behind zero-shot segmentation is to transfer knowledge from seen classes (classes seen during training) to unseen classes (classes not seen during training) by utilizing shared attributes or semantic embeddings. By understanding the relationships between different classes and leveraging this information, the model can generalize to segment objects or regions belonging to unseen classes.

Segment Anything Model (SAM) is a zero-shot segment framework, developed by the company Meta Platforms, Inc. (California, United States) to build a starting point for foundation models for image segmentation. It was introduced by Kirillov et al. (Kirillov et al., 2023). This takes inspiration from the field of NLP (Natural Language Processsing) where foundation models and large datasets (worth billions of tokens) have become commonplace. The project leads to the creation of a large dataset, a segmentation model, and is fed back into the loop. The final dataset includes more than 1.1 billion segmentation masks collected on 11 million licensed and privacy preserving images. It should be emphasized that this framework is still under development and is optimized to work with segmentable images, i.e., images that have good contrast and are easily differentiated. But in the future this tool can be specialized to focus on different environments, such as medical.

Semiautomatic segmentation of medical images has served as a valuable intermediate step between manual and automatic approaches. It has allowed for user interaction and guidance to refine segmentation results while reducing the manual effort required. However, the field is continuously advancing towards fully automatic segmentation methods. The goal is to minimize user involvement and rely on advanced computational algorithms, such as deep learning, for accurate and efficient segmentation of medical images.

2.3. Automatic segmentation

Automatic segmentation of medical images involves techniques such as machine learning, deep learning and classical pre- and post-processing of images to automatically analyze and identify the desired structures. The goal is to achieve accurate and efficient segmentation results, reducing the need for manual effort and minimizing subjectivity.

Automatic models have been developed to identify and properly segment the different regions and subregions of the mouse brain. Mesejo et al. (Mesejo et al., 2012) developed a two-step automated segmentation method for the hippocampus in histological images (5). Initially, they maximize the overlap of a parametric deformable model with two important reference substructures in the brain image using differential evolution. This step guides the determination of the region of interest. In the second step, a thresholding technique based on Otsu's method is applied to the points identified in the previous step. Finally, Random Forest is used to extend the segmentation to regions not covered by the model. The method achieved an average segmentation accuracy of 92.25% and 92.11% on independent test sets composed of 15 real and 15 synthetic images, respectively.

Several frameworks have been developed for an automatic segmentation of mouse brain regions, mostly for MRI and ultrasound imaging, but none for histological imaging. Until recently, Barzekar et al. (Barzekar et al., 2023) provide a model capable of efficiently detecting two subregions on histological slides, Substantia Reticular part (SNr) and Substantia Nigra Compacta, dorsal tier (SNCD) in all images, with a U-Net-based architecture. They compare the performance of their model with various combinations of encoders, image sizes, and sample selection techniques. In addition, to increase the sample set they opted for data augmentation, which provided data diversity and robust learning. The model was trained on approximately one thousand annotated 2D coronal brain images stained with Nissl/Hematoxylin and the enzyme Tyrosine Hydroxylase (TH, an indicator of dopaminergic neuron viability). The final reach a

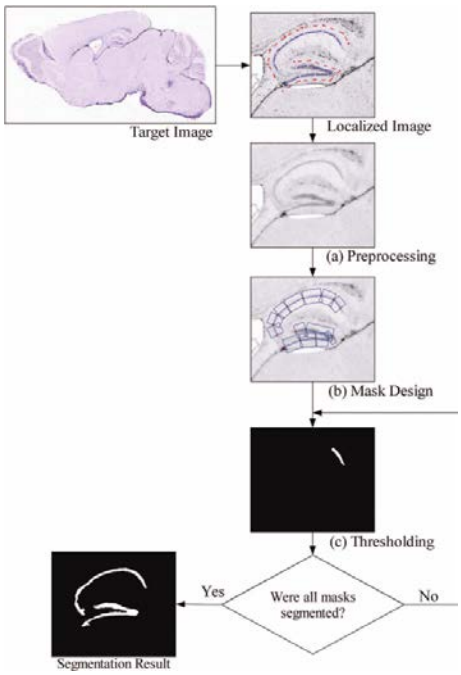


Figure 5: Automated Hippocampus segmentation pipeline propose in Mesejo et al. (Mesejo et al., 2012)

Dice coefficient of 87% for the task.

10 years separate the methods proposed by Mesejo et al. (Mesejo et al., 2012) and by Barzekar et al. (Barzekar et al., 2023). Incredibly, over this period of time, no models have been developed or proposed that are able to work with this type of imaging of the mouse brain. As mentioned above, most of the proposed methods are applied for MRI. Furthermore, the literature review revealed the lack of automatic systems for the detection of different areas of the mouse brain specifically for working with histological images as input data. One of its main causes is the lack of properly labeled data which, in some cases, is carried out by people without the necessary knowledge for proper labeling. Furthermore, histological images have high resolution (e.g. 28,000 pixels x 14,000 pixels), leading to a lack of adequate hardware for training and testing models, given the size of the images exceeding 1GB each. This complicates the management of the images when a deep learning model is to be trained and used.

3. Material and methods

I divided the automatic segmentation for the different regions of the mouse brain into three main parts: Dataset preparation, deep learning model and image post-processing.

3.1. Dataset preparation

The NeuroGEMM laboratory (University of Burgundy, Dijon) has several thousands of 2D images of murine brain samples manually curated and segmented. The overall process that the laboratory uses to analyze a large number of mouse brains, which is robust and simple, is detailed below. Four examples of images are shown in Figure 6.

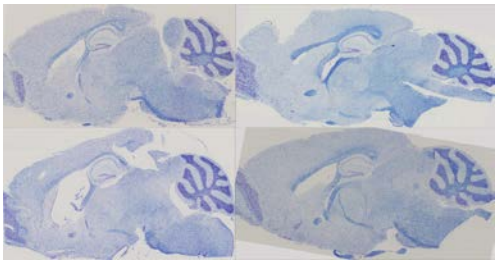


Figure 6: Examples of images of the mouse brain in sagittal view.

Brains are collected from 14 to 16-week-old KO mice, fixed in paraformaldehyde and embedded in paraffin. Sections of 5 μm are cut with a microtome using three coronal or one sagittal plane depending on the project. The section of interest is referred to as the "critical section". Sections are deparaffinized and stained with fast luxol blue and cresyl violet. Luxol is a blue stain that stains the myelin revealing the axons. Cresyl stains the Nissl bodies present in the rough endoplasmic reticulum of neurons, thus showing the cell body. Finally, histological slides are scanned with a high-resolution scanner (NanoZoomer 2.0HT, Hamamatsu, Japan) to obtain digitized slides at high enough resolution to see every single cell (around 20,000 pixels by 12,000 pixels with a resolution of 0.455 micrometer per pixel). Prior to data anlysis, a "quality control" is performed to establish the degree of variation in critical section distance to the target stereotaxic plane, followed by an evaluation of symmetry, staining and image quality. Then, the next step is to measure the areas and length of the defined brain parameters/regions following the standard operating procedures (SOPs) developed by the laboratory. Fiji software is used to measure both surface area and distances at each slice given by landmark annotations for each region of interest. Both measures are manually annotated and saved in ROI format, which is a file type that stores information (landmarks) about a specific region of an image that the user intends to analyze separately from the rest of the image. Figure 7 shows an example of visualization of the landmarks stored in the cerebellum annotation. Since the overarching goal of the data analysis is to identify morphogenes, all the analysis is done genotype-blind until this step.

Following analysis, the data are also checked for human error or outliers. The laboratory has an automatic program that calculates outliers within the interquartile range (IQR) of 1.2. Most of the time, outliers are due to asymmetries or suboptimal coordinates. In addition to checking if all regions, or part of all, should be taken, correct labeling is evaluated. In our project, we worked

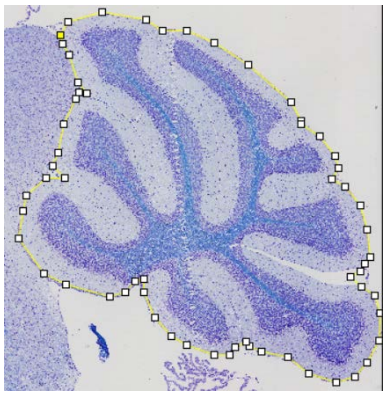


Figure 7: Landmarks, manually taken, for the Cerebellum (TC).

with the sagittal view and 24 regions of interest listed in Table 1 and in Figure 8.

Table 1: Neuroanatomical Features

TAG	Full Name
aca	anterior commisure
cc	corpus callosum
f	fornix
fi	fimbria
fp	fibers of the pons
och	optic chiasm
sm	stria medularis
TB	Total Brain
TCTX	Total Cortical area
TC	Total Cerebellum
IGL	Intra Granular Layer
LV	Lateral Ventricle
TTh	Thalamus
CPu	Caudate Putamen
HP	Hippocampus
TILpy	pyramidal cell layer
DG	Dentate Gyrus
Pn	Pontine nucleus
SN	Substancia Nigra
Cg	Cingulate cortex
DS	Dorsal Subiculum
InfC	Inferior Colliculus
SupC	Superior Colliculus
VMHvl	Ventro Median Hypothalamus ventro lateral part

The work began with the preparation of the dataset that was used to train the model.

Due to the number of regions and the diversity of approaches that could be implemented for each one, it was chosen to work with a general model for all regions. Due to the large size of the image files, a minimum workable resolution was evaluated as a trade-off between calculation time for training and annotation accuracy.

3.1.1. Landmarks to binary masks

The database was crossed-check for existence of both images and .roi files.

Once this was verified, we checked if images and landmarks defined by the .roi file overlaid properly and matched in size. Regions were binarized and their size was reduced from high resolution to two lower resolutions, 512x256 and 2048x1024 to fit the medium scaled images. A bilinear interpolation algorithm was used for this purpose. It is a method used to estimate values between two known values in a grid or image. It calculates the intermediate value based on a weighted average of the surrounding four pixels. Subsequently, masks were saved and some random checks were done to visually verify that no regions were saved with a different name (e.g. saved the fimbria instead of the anterior commisure area). Based on the binary masks created for all the regions, the location of each of them was analyzed as a fourth step of data revision. It was found that several regions overlapped due to manual annotation of their contours. Thus, a method for the individual segmentation of the each region of interest individually was proposed instead of a multi-class approach.

3.1.2. Brain division

The first step was to divide brain regions of interest into two groups. The first group consists of regions where the entire image are used as input to train the model. Within this group are: InfC, SupC, IGL, TC, TC, SN, Pn, fp and TB. Meanwhile, for the second group, the total brain area (TB) was used as the working boundary area and the regions within this area were localized. This group consists of: aca, cc, Cg, CPu, DG, DS, f, fi, HP, IGL, LV, och, sm, TCTX, TTH and VMHvl. Figure 9 shows the grouping of the regions.

3.2. Deep learning models

We tested several learning models. Initially, we chose to work and test these models at low resolution (512x256) to verify their ability to capture important features in the mouse brain.

The first model, U-Net, proposed by Ronneberger et al. (Ronneberger et al., 2015), was used as a starting point to segment brain regions, since this architecture is often used in medical image segmentation. We chose to work with a depth of 5 levels and with feature maps in the encoder of 3, 16, 32, 64, 128 and for the decoder 256, 128, 64, 32 and 16, being the initial configuration of the architecture.

Each experiment was trained for 100 epochs with a minibatch size of 32 images. An Adam optimization was performed and weights were saved for the epoch that produced the last validation loss. A starting learning rate was fixed at 0.01 with a scheduler function that reduces the learning rate by a factor of 10 after 15 consecutive epochs. An early stopping function was im-

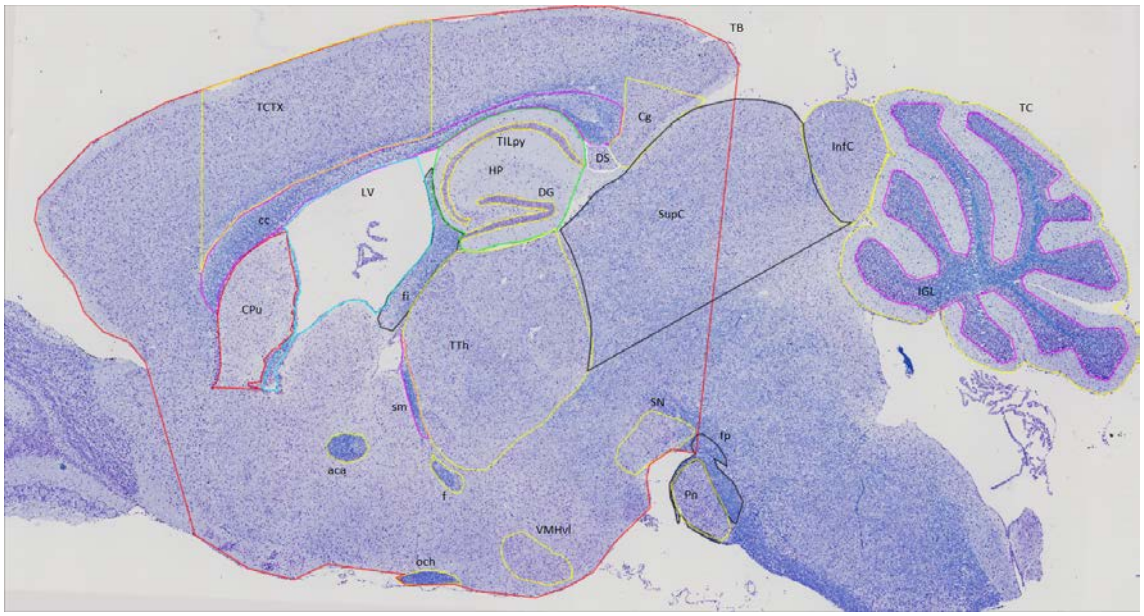


Figure 8: Regions of interest listed in the Table 1, taken by human user, within the mouse brain in sagittal view.

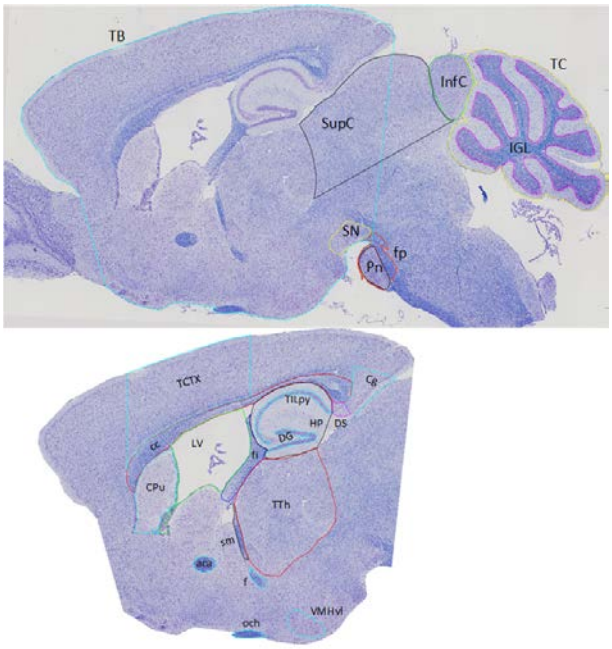


Figure 9: First (top) and second (bottom) group of regions

plemented to make the model training process time efficient. The model was trained with the Pytorch framework using a NVIDIA A100 GPU. Two different loss functions were tested, binary cross entropy with logits loss (BCE) and Dice loss. The Dice loss function is defined as follows:

$$\text{Dice Loss} = 1 - \frac{2\sum_{i=1}^{N} p_i \cdot t_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} t_i^2}$$

Where:

N : Total number of pixels
p_i : Predicted probability/label for pixel
t_i : Ground truth label for pixel

The binary cross entropy with logits loss function is defined as follows:

$$\text{BCE with Logits Loss} =$$

$$\frac{1}{N}\sum_{i=1}^{N}(\log(1+\exp(-t_i \cdot p_i)) + \max(0, p_i) - t_i \cdot p_i)$$

Where:

N : Total number of samples
p_i : Predicted logits for sample
t_i : Ground truth labels for sample

The Dice loss function was chosen because it offers better metrics and ability to describe model performance. The results obtained did not meet the objective in terms of accuracy. Therefore, the need for a change in the architecture became evident (see results sections)

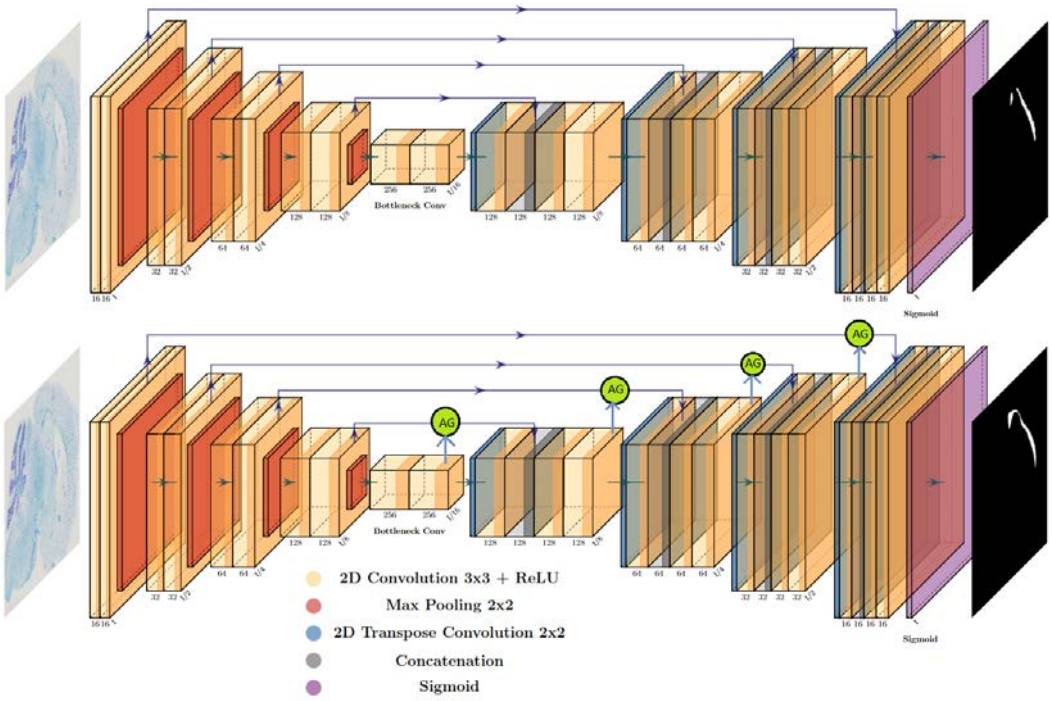


Figure 10: Comparison between U-Net (top) vs Attention U-Net (bottom), architectures for training a single region.

as it required significant post-processing of the image to fill holes in the binary masks by using morphological operations and Dice loss scores were relatively poor.

The second model is based on Okay et al. (Oktay et al., 2018) who proposed an architecture called Attention U-Net which adds attention blocks, which dynamically weight the importance of different image regions during the segmentation process. This enables the network to focus on relevant features and enhance the accuracy of the segmentation results. The same depth and feature maps as the previous architecture were maintained. In Figure 10 is presented a visual comparison between both architectures.

With better segmentation produced we decided to modify the features maps to improve segmentation precision, and used an encoder/decoder feature maps consisting of: 3, 64, 128, 256, 512 and 1024, 512, 256, 128 and 64, respectively.

A third test was carried out using the framework implemented by Isensee et al. (Isensee et al., 2021), named nnU-Net, for medical image segmentation. It is an extension of the U-Net architecture and is specifically designed for medical imaging applications. nnU-Net provides a standardized and reproducible pipeline for training and evaluating segmentation models on various medical image datasets.

At the end of this round, using low definition images, it was concluded that the U-Net architecture was not able to acquire enough information to adequately delimit the regions of interest. Indeed, it was necessary to apply extenssive post-processing of the images to improve segmentation. This model was thus discarded from the next round of testing which used medium resolution (2048x1024).

For the Attention U-Net, the number of levels had to be increased from 5 to 7 in order to obtain consistent segmentation results. The feature maps are as follows: encoder 3, 64, 126, 256, 512, 1024, 2048 and decoder: 4096, 2048, 1024, 512, 256, 128, 64.

In Figure 11 is shown a standard workflow for all the regions within the mouse brain.

In order to train and test the results of the models, the database of all masks and images was divided into 3 groups. We used 70, 15 and 15 for training, validation and testing, respectively. Once the training was finished, performances of the models were tested in the different regions and for the different resolutions previously established.

3.3. Image processing

After training the deep learning models for the two resolutions, it was necessary to return to the original size of the images, so a bicubic interpolation algorithm was used instead of the previously used bilateral one. This interpolation calculates the intermediate value based on a weighted average of the surrounding 16 pixels. This technique is commonly used in image processing to resize or rescale images while maintaining greater sharpness and detail compared to bilinear interpolation.This change was necessary because the main objective was to segment with a high level of accuracy the regions of

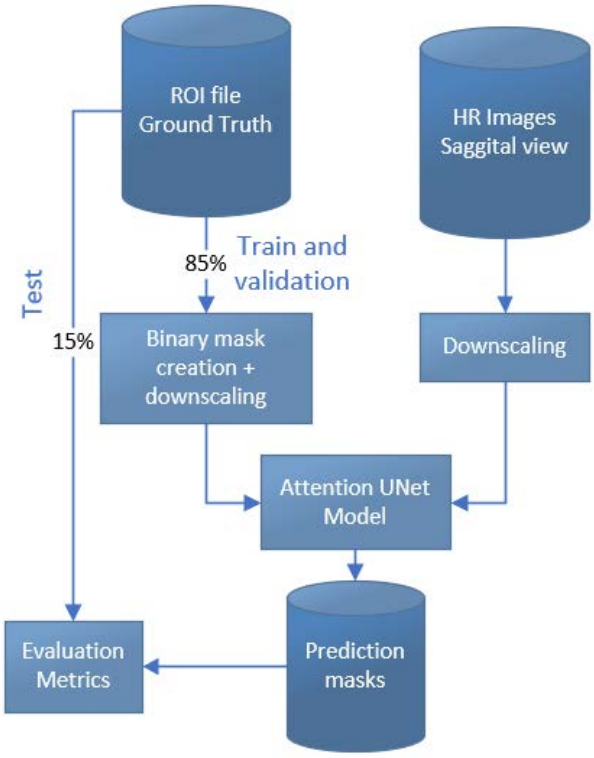


Figure 11: Workflow proposed to evaluate the accuracy of the training. This example is for Attention U-Net in medium resolution 2048x1024

interest within the mouse brain. Therefore, bicubic interpolation helps to convert the finer details of the mask. Unfortunately, even if you have the best interpolation method, irregularities in the fine definition of the contour are encountered. Figure 12 shows an example of the jagged edges present as a result of the process of returning the mask to its original size.

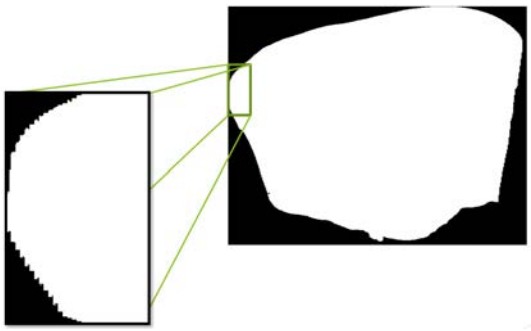


Figure 12: Example of jagged contours

The Douglas-Peucker algorithm was used to solve this problem. It is a method for simplifying polylines or curves by reducing the number of points while preserving their shape. It selects significant points that contribute to the overall shape and eliminates less significant ones (Mokrzycki and M, 2012). By iteratively calculating distances and selecting the point with the maximum distance from a line segment, the algorithm removes redundant points and simplifies the curve while retaining its essential characteristics. Figure 13 shows in a visual way the result of its application. Several tests were performed to achieve a balance between the number of points delivered by the polygon contour from the binary mask, and the accuracy in annotating the contour of the region of interest.

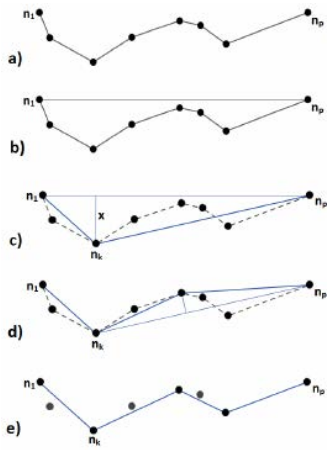


Figure 13: Example of the Ramer–Douglas–Peucker algorithm. a) Input curve, b) - d) specified stages of the Ramer–Douglas–Peucker algorithm, e) output curve with reduced number of points (Mokrzycki and M, 2012).

As a final result of our proposal, we have a set of landmarks for each of the 24 regions of interest within the mouse brain for histopathological imaging in sagittal view. Figure 14 shows the final pipeline of the model. The results of each of the phases and the models used are presented in section 4.

3.4. Evaluation metrics

The following metrics listed below will be used to evaluate the performance of the models worked on:

- Dice coefficient or Dice similarity coefficient, is a metric commonly used to evaluate the accuracy of segmentation results. It measures the overlap between the predicted segmentation and the ground truth by calculating the ratio of twice the intersection of the two regions to the sum of their sizes.

$$Dice\ coefficient = \frac{2 * Intersection}{Prediction + GroundTruth}$$

- False Positive Rate (FPR) is a metric that measures the proportion of incorrect positive predictions made by the model. A lower false positive rate indicates better performance, as it indicates a lower rate of false alarms or incorrect positive predictions.

$$FPR = \frac{FP}{FP + TN}$$

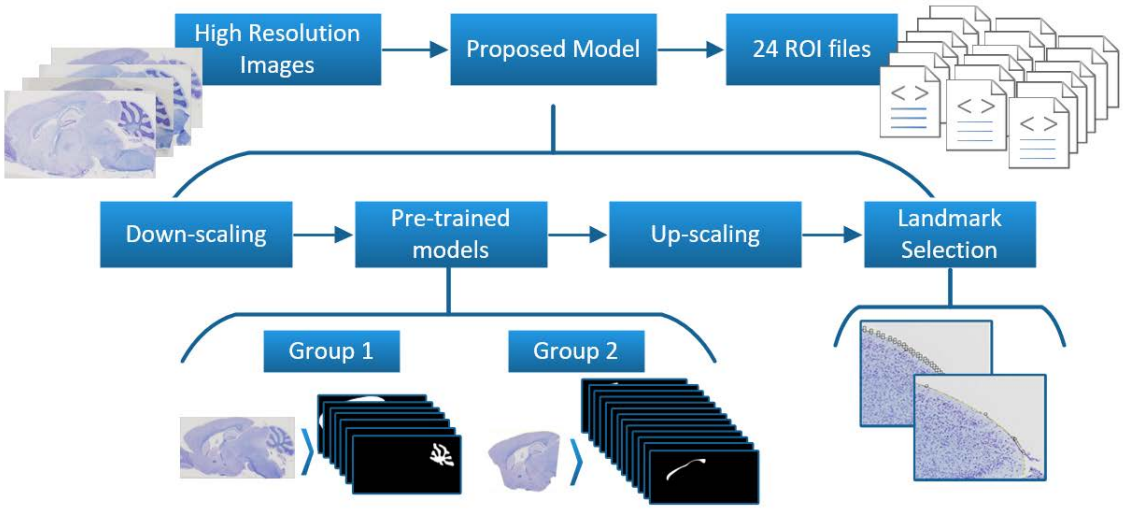


Figure 14: Final pipeline of the proposed method

FP = False Positive, TN = True Negative

- False Negative Rate (FNR), measures the proportion of missed positive predictions by the model. A lower false negative rate is desired as it signifies a lower rate of missed detections or incorrectly classified negatives, indicating better sensitivity and accuracy in capturing the target structure or region.

$$FNR = \frac{FN}{FN + TP}$$

TP = True positive, FN = False Negative

Both FPR and FNR will be used to evaluate the response of the models at pixel level.

- Hausdorff Distance (HD) measures the dissimilarity between two sets of points or contours. It quantifies the maximum distance between any point in one set to the closest point in the other set.

$$HD(A, B) = max(max(d(a, B)), max(d(b, A)))$$

where:
d(a, B) represents the minimum distance between a point a in set A and the closest point in set B.
d(b, A) represents the minimum distance between a point b in set B and the closest point in set A.

- Shapiro-walk test is a method to evaluate if a measure follows a normal distribution. This measure will help decide what other approaches will be used to evaluate the model.

- The Student t-test is a statistical test used to determine if there is a significant difference between the means of two groups which follows a normal distribution. It compares the means while considering variability within each group and sample size.

- Wilcoxon signed-rank test is a nonparametric statistical hypothesis test used to compare the location of two populations from two paired samples. It is a paired differences test like the Student t-test, but can be used with data that do not follow a normal distribution.

- Bland-Atman plot is a visual method used to analyze the agreement between two different assays. The objective is to determine whether there is a systematic bias or significant variability between the two methods. (Bland and Altman, 1986).

3.5. Segment Anything Model

As a test, the recently launched automatic segmentation tool SAM (Kirillov et al., 2023), was used. This tool was used to test its efficiency in segmenting histological images, in this case of the mouse brain. Our model masks were used as input to help localize, partially, the different regions of the brain. The results of the different tests are shown in the next section.

4. Results

This section presents the results of all the methods and techniques applied for the realization of the proposed model.

4.1. Data preparation

After the whole process of preparing the database, and going through the different stages of review. As a result, we obtain the following area masks to work with each region of interest within the mouse brain (Figure 15).

It should be noted that there is a different number of masks per region. This difference in quantity is due to several factors such the loss of information over the

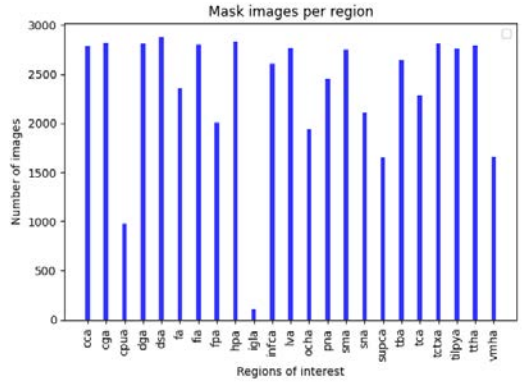


Figure 15: Number of mask images per region of interest area.

years, mislabeling of the areas or because the region is not present in the brain slice. A notorious example is with the IGL area, which due to human errors when taking the measurement could be located in another area of the brain slice. This region is left uncorrected to check the number of images required for a correct segmentation, given a large region with a good contrast to the others.

Once the binary image dataset is created, for each of the regions, we proceed to the deep learning training stage.

4.2. Deep Learning

In order to start with the training of the models, the first step was the selection of the loss function. The best results were obtained with the use of a Dice loss function. A comparison between the two revised loss functions is presented in Figure 16, where the Dice similarity coefficient evaluation metric was used. The architecture used for this purpose is the normal U-Net.

Once the loss function has been selected, the different architectures proposed in the methods section can be trained.

4.2.1. U-Net 5 levels

The results of training with the U-Net architecture are presented for three different regions of the mouse brain in Figure 17. The input images have the size of 512 pixels by 256 pixels.

Morphological operations such as dilation and erosion were applied (Figure 18) to fill in incomplete areas and in some cases to eliminate erroneously segmented pixels.

Since there is a diversity in the results due to the characteristics of the images, a post-processing stage cannot be generalized to correct for the absence or presence of additional pixels. Therefore, this stage showed a decrease in the Dice value as in the FPR.

Due to the lack of accuracy in image segmentation with the basic U-Net architecture, it was decided to

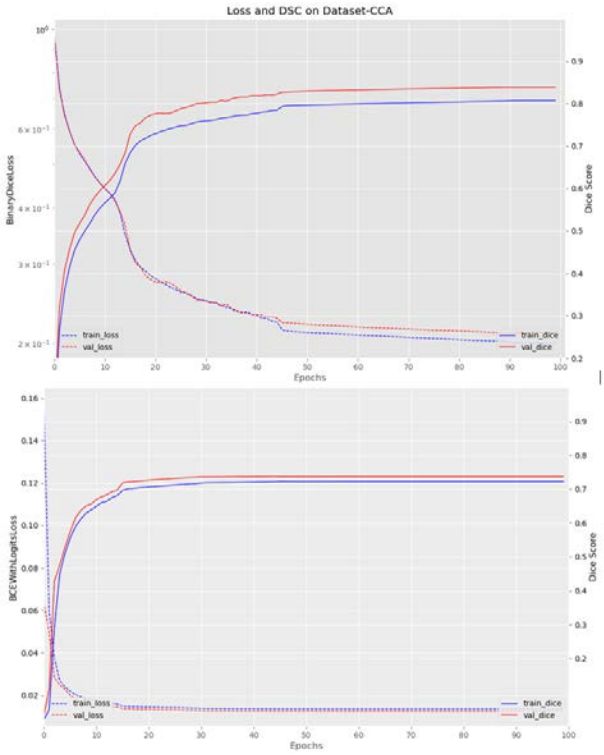


Figure 16: Comparison performance while training models for the corpus callosum (cc) using Dice loss versus binary cross entropy with logits loss with Dice similarity coefficient as evaluation metric.

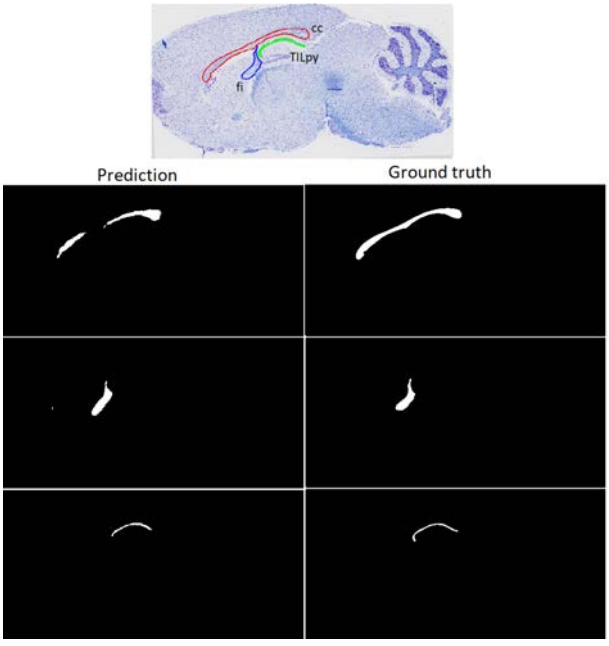


Figure 17: From left to right: prediction of the regions and ground truth masks for (top) corpus callosum, (middle) fimbria, (bottom) pyramidal cell layer TILpy.

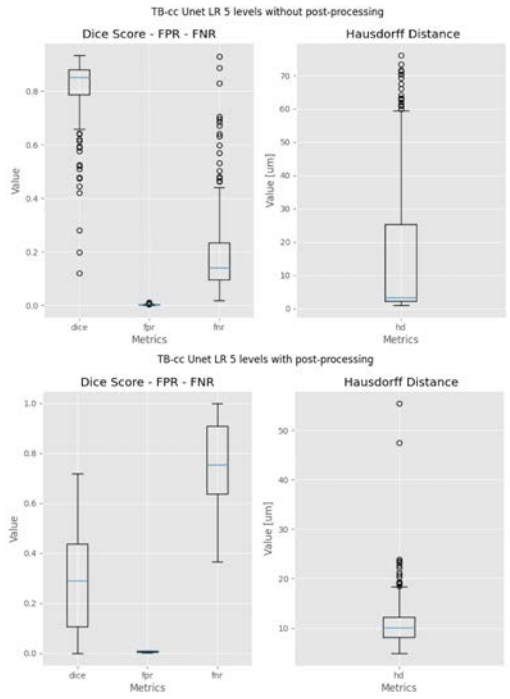


Figure 18: Results before and after applying morphological operations for the segmentation of the corpus callosum.

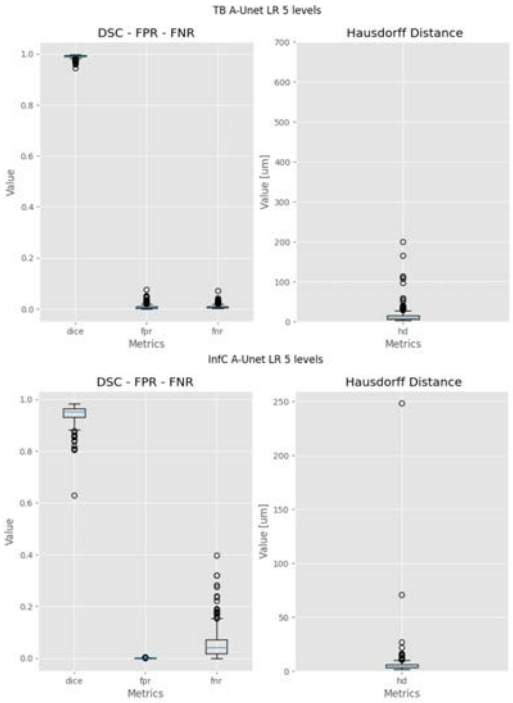


Figure 20: Results for TB (top) and InfC (bottom) while training with low resolutions (521x256) images with a 5 levels Attention U-Net

switch to an U-Net variant that includes attention gates, Attention U-Net, to guide the segmentation.

4.2.2. Attention U-Net 5 levels

The results of training with the Attention U-Net architecture are presented for two different regions of the mouse brain (TB and InfC) in Figure 19. The input images have the size of 512 pixels by 256 pixels.

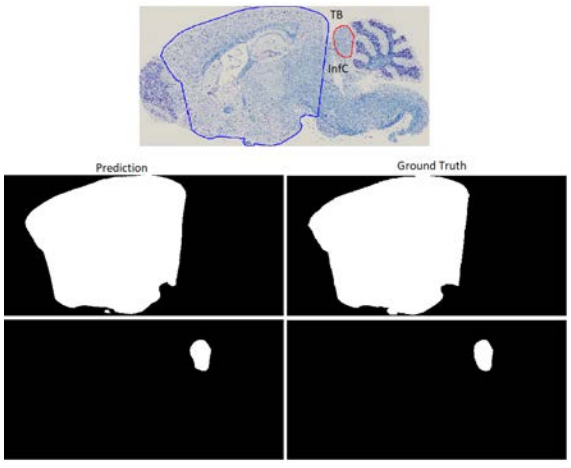


Figure 19: From left to right: prediction of the regions and ground truth masks for (top) total brain area TB and (bottom) inferior colliculus InfC.

In Figure 20 is shown the output evaluations metrics for the training of two particular regions, TB and InfC.

What is being tested now is whether segmentation accuracy can be maintained, or better, with higher resolution images. For this purpose, the same test carried out for a resolution of 512x256 is performed, with the increase to 2048x1024 in resolution for the input image. The training response, for the same regions as above is presented in the following Figure 21.

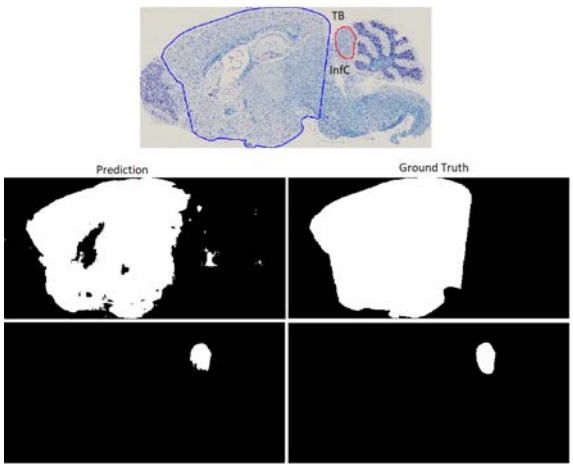


Figure 21: From left to right: prediction of the regions and ground truth masks for (top) total brain area TB and (bottom) inferior colliculus InfC.

In Figure 22 is shown the output evaluations metrics for the training of TB and InfC with the new input resolution images.

The results obtained clearly showed the need to

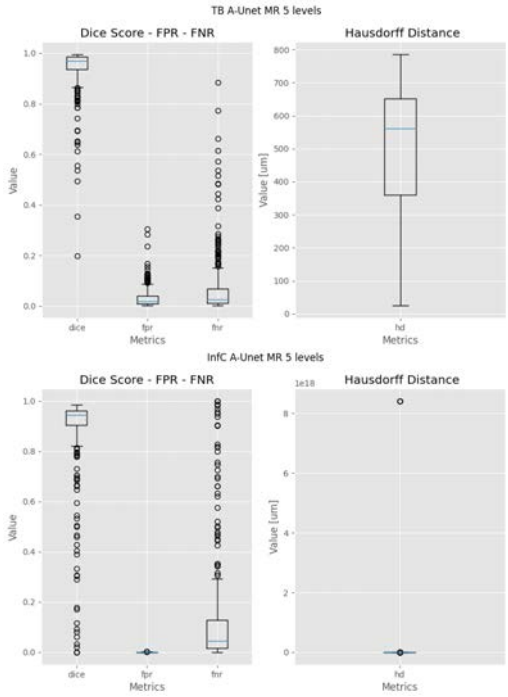


Figure 22: Results for TB (top) and InfC (bottom) while training with the second resolutions images with a 5 levels Attention U-Net

change the way in which the features are obtained within the model, but without losing the details at a smaller scale. It was decided to increase the depth of the architecture, so that the model could obtain more features and improve its performance.

4.2.3. Attention U-Net 7 levels

The results of training with the Attention U-Net architecture with 7 depth levels are presented for the same mouse brain regions previously worked in Figure 23.

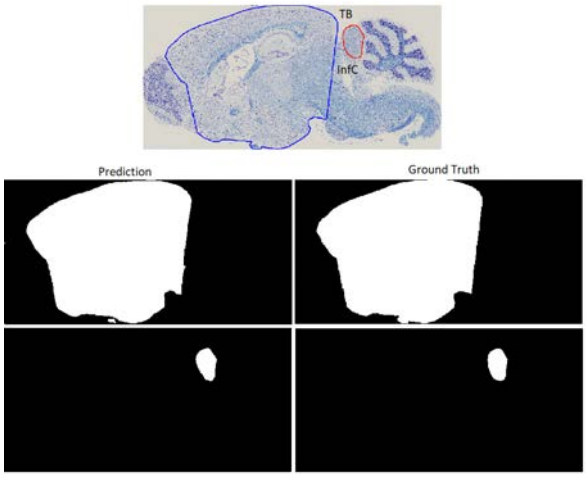


Figure 23: From the left to the right:predictions for (top) total brain area TB and (bottom) inferior colliculus InfC with a 7 levels depht Attention U-Net.

In Figure 24 is shown the output evaluations metrics for the training of TB and InfC.

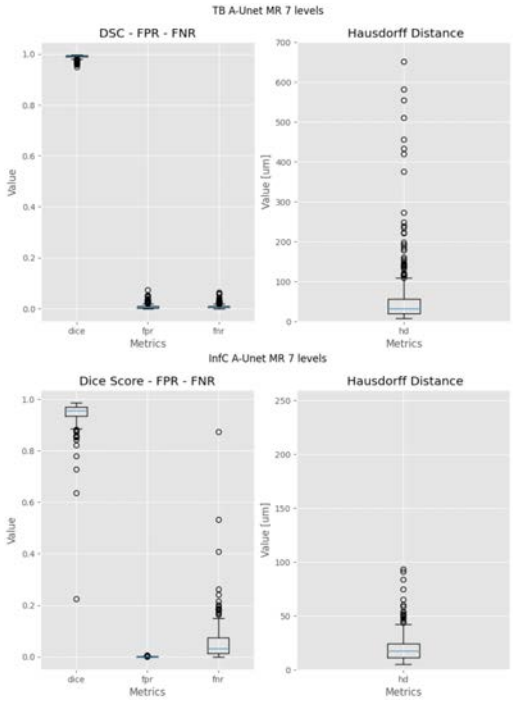


Figure 24: Results for TB (top) and InfC (bottom) while training with the second resolutions images with a 7 levels Attention U-Net.

4.3. Image post-processing

After the selection of the models, the masks are post-processed. They are converted from binary images to point vectors so that they can be visualized using Fiji/ImageJ software. Figure 25 shows an example of contour selection and curve approximation. The vectors are saved in ROI format. Figure 26 compares the post-processing performance with the two different resolutions (512x256 and 2048x1028).

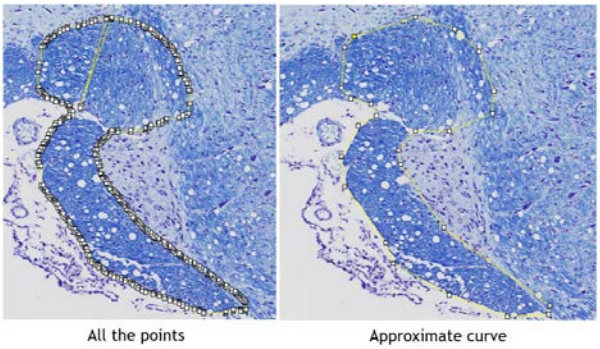


Figure 25: Comparison between before and after the curve approximation for the fp (fibers of the pons).

4.4. Quantitative analysis

The performance of the different models with the best responses, Attention U-Net of 5 levels with 512x256

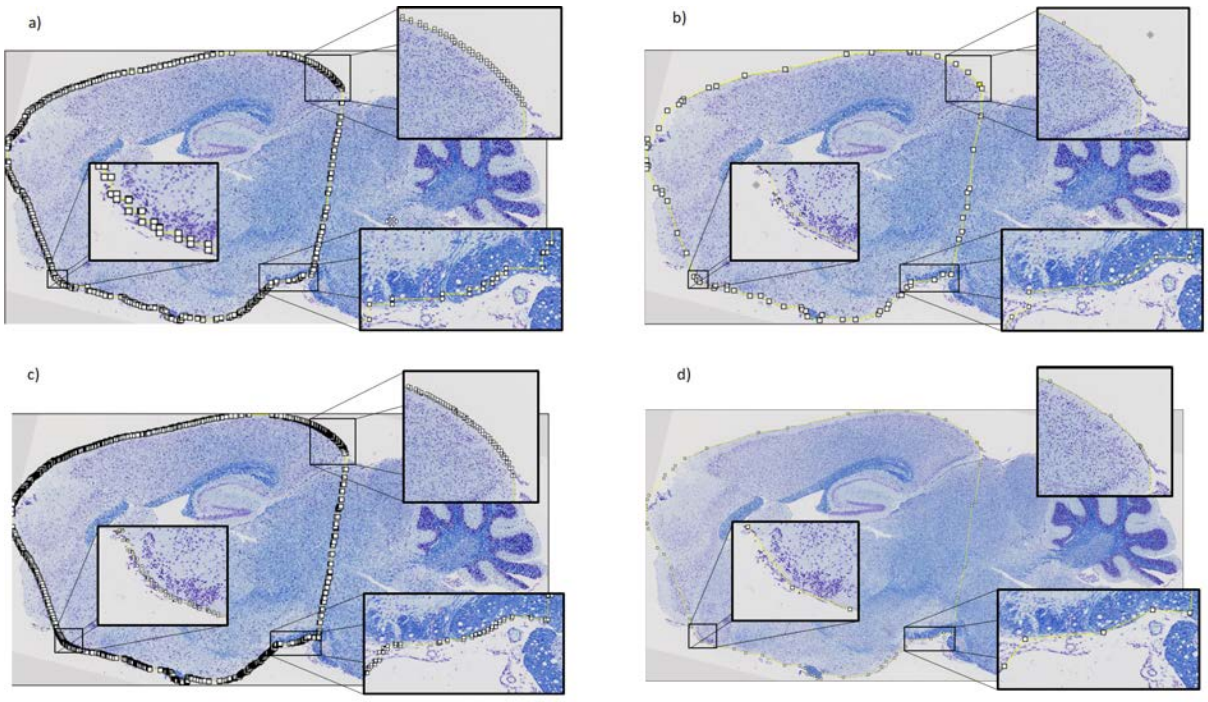


Figure 26: Example of converting contours to landmarks for the total brain area a) original number of points (805) in 512x256 resolution b) output landmarks (76) with resolution images in b) c) original number of points (1501) with 2048x1024 resolution images d) final landmarks (56) with resolution images in c).

Table 2: Evaluation metrics after image post-processing with an Attention U-Net 5 levels 512x256 resolution images

TAG	DSC	STD	FPR	STD	FNR	STD	HD μm	STD μm
aca	0.9617	0.0403	**0.0001**	0.0001	0.0389	0.0539	**0.5812**	0.5522
cc	0.9384	0.0634	0.0008	0.0005	0.0599	0.0718	1.7735	4.5100
f	0.8687	0.0901	**0.0001**	0.0001	0.1289	0.1257	1.3862	1.6350
fi	0.9155	0.0556	0.0005	0.0004	0.0848	0.0770	2.9190	1.9320
fp	0.7208	0.1681	0.0010	0.0013	0.2630	0.1995	7.6976	6.6295
och	0.9027	0.0938	**0.0001**	0.0001	0.0972	0.1119	1.4469	1.7490
sm	0.8729	0.1037	0.0003	0.0005	0.1104	0.1230	3.9239	5.9736
TB	**0.9911**	0.0058	***0.0077***	0.0084	**0.0085**	0.0072	5.3201	4.5393
TCTX	0.9789	0.0129	0.0010	0.0008	0.0213	0.0208	2.0512	1.6605
TC	0.9888	0.0078	0.0010	0.0006	0.0119	0.0134	2.8932	2.7347
IGL	0.9228	0.0474	0.0071	0.0042	0.0072	0.0057	2.4526	1.2877
LV	0.7361	0.2533	0.0032	0.0039	0.1987	0.2590	***11.0957***	7.2475
TTh	0.9520	0.0244	0.0022	0.0019	0.0453	0.0371	3.7287	1.7094
CPu	***0.6594***	0.3065	0.0006	0.0008	***0.3306***	0.3258	5.6841	4.8462
HP	0.9828	0.0065	0.0004	0.0002	0.0171	0.0102	1.3799	0.7973
TILpy	0.8886	0.0332	0.0002	0.0001	0.1056	0.0593	1.7654	1.3216
DG	0.9238	0.0242	0.0002	0.0001	0.0682	0.0443	0.9397	0.7214
Pn	0.9418	0.0718	0.0002	0.0002	0.0576	0.0845	1.1540	0.9904
SN	0.7496	0.2144	0.0007	0.0007	0.2072	0.2323	3.2851	2.3954
Cg	0.9098	0.0450	0.0007	0.0006	0.0817	0.0671	2.5104	1.3651
DS	0.8712	0.0538	**0.0001**	0.0001	0.1333	0.0887	1.2464	0.5641
InfC	0.9424	0.0355	0.0006	0.0006	0.0579	0.0571	2.2038	1.3951
SupC	0.9483	0.0249	0.0035	0.0032	0.0479	0.0364	5.3587	2.7498
VMHvl	0.7660	0.1712	0.0005	0.0005	0.2259	0.2195	3.0721	2.7958

Table 3: Evaluation metrics after image post-processing with an Attention U-Net 7 levels 2048x1024 resolution images

TAG	DSC	STD	FPR	STD	FNR	STD	HD μm	STD μm
aca	0.9660	0.0394	**0.0001**	0.0001	0.0276	0.0573	**5.2889**	11.0271
cc	0.9365	0.0347	0.0012	0.0007	0.0421	0.0463	15.1756	14.8965
f	0.8828	0.0957	**0.0001**	0.0002	0.0822	0.1161	12.6860	40.1613
fi	0.9179	0.0810	0.0004	0.0004	0.0798	0.1035	27.9981	28.0597
fp	***0.7047***	0.1758	0.0013	0.0015	***0.2383***	0.2184	***66.8368***	52.7969
och	0.9214	0.0801	**0.0001**	0.0002	0.0613	0.1021	12.8222	22.2288
sm	0.8775	0.0927	0.0003	0.0004	0.0997	0.1256	31.4965	52.4748
TB	**0.9914**	0.0062	***0.0079***	0.0085	**0.0078**	0.0080	42.9213	42.1242
TCTX	0.9780	0.0280	0.0010	0.0008	0.0205	0.0387	20.3151	28.2891
TC	0.9902	0.0053	0.0010	0.0006	0.0087	0.0083	24.1482	23.9261
IGL	0.9267	0.0865	0.0036	0.0042	0.0546	0.1240	33.7096	43.5162
LV	0.9452	0.1069	0.0005	0.0010	0.0493	0.1058	41.9781	64.0020
TTh	0.9515	0.0268	0.0021	0.0017	0.0492	0.0435	34.1868	16.4720
CPu	0.7918	0.2614	0.0010	0.0013	0.1846	0.2654	43.7350	42.5232
HP	0.9848	0.0068	0.0004	0.0002	0.0122	0.0109	12.6344	7.6397
TILpy	0.8852	0.0812	0.0003	0.0001	0.0814	0.1050	18.8284	24.7633
DG	0.9302	0.0770	0.0002	0.0002	0.0465	0.0862	8.0014	11.5809
Pn	0.9500	0.0600	0.0002	0.0002	0.0391	0.0667	9.2620	8.7277
SN	0.7647	0.1958	0.0006	0.0007	0.2014	0.2184	28.2791	20.1335
Cg	0.9087	0.0612	0.0007	0.0006	0.0817	0.0873	23.3410	19.6141
DS	0.8842	0.0784	**0.0001**	0.0001	0.1012	0.1103	10.8247	6.6032
InfC	0.9442	0.0613	0.0006	0.0005	0.0542	0.0784	19.6080	13.8624
SupC	0.9483	0.0249	0.0035	0.0032	0.0479	0.0364	11.7670	6.0383
VMHvl	0.8039	0.1754	0.0005	0.0004	0.1797	0.2095	23.7869	23.7924

image resolution and Attention U-Net of 7 levels with 2048x1024 image resolution, were evaluated to start the data analysis (Tables 2 and 3). The best values are highlighted in bold and the worst values are emphasized in bold italics. With the first model, the best result was TB (total brain area) and the worst was CPu (caudade putamen) for DSC with 99.11% and 65.94% respectively. Meanwhile, for the second model, TB remains the best value with 99.14% and now the worst value is fp (fibers of the pons) with 70.47%, both for DSC.

The Saphiro-walk test was applied to prove the hypothesis that the difference between the predicted and true region masks follows a normal distribution. The test led to reject the hypothesis of normality of the data with 95% confidence for all regions of interest.

Subsequently, the nonparametric Wilcoxon test was chosen to evaluate the relationships between the two DSC values because this test allows working with samples that do not fulfill a normal distribution, in addition to allowing working with data that are smaller in size. The Wilcoxon test was performed at the different resolutions and changes in the architecture.

Some examples of Bland Altman plots are presented in Figure 27, specially for some areas which were difficult to segment.

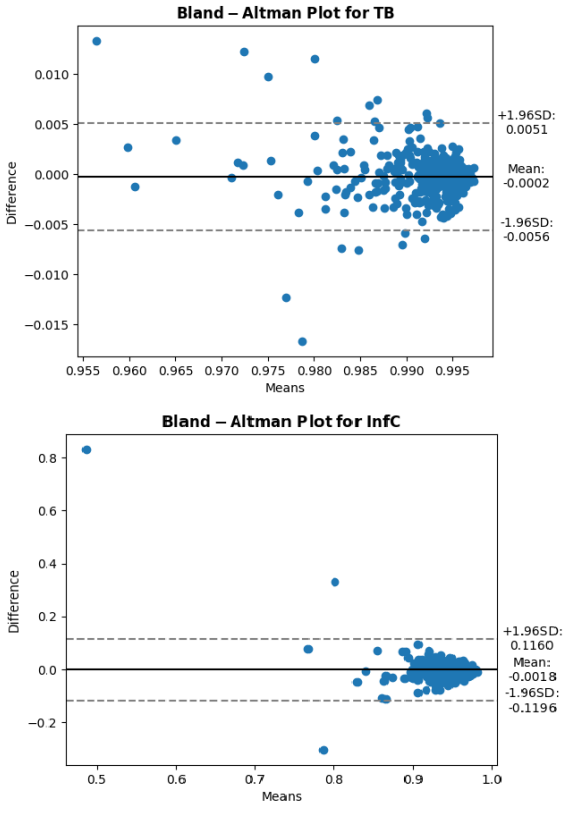


Figure 27: Comparison between DSC with both resolutions for TB and InfC

Table 4: nnU-net framework evaluation metrics

TAG	DSC	STD	HD	STD
fp	0.7605	0.1490	78.9387	100.6994
CPu	0.7701	0.2910	66.6696	162.7281

Table 5: nnU-net framework evaluation metrics

TAG	FPR	STD	FNR	STD
fp	0.0011	0.0011	0.1525	0.1337
CPu	0.0012	0.0025	0.1963	0.3071

4.5. nnU-Net

In addition, the segmentation was checked with the nnU-Net framework, in which the respective metrics were performed, where the lowest DSC values were obtained with our models. The results obtained for fp and CPu are presented in the Tables 4 and 5.

In the end, the results obtained by our proposed method are at the level of the nnUnet framework. In one region we obtained better results in both DSC and HD, while in another region we obtained worse results.

4.6. SAM

Finally, the performance of the new zero-shot mode SAM segmentation was tested for mouse brain histological samples (Figure 28) in three specific regions: anterior commisure area (aca) and Ventro Median Hypothalamus ventro lateral part (VMHvl).

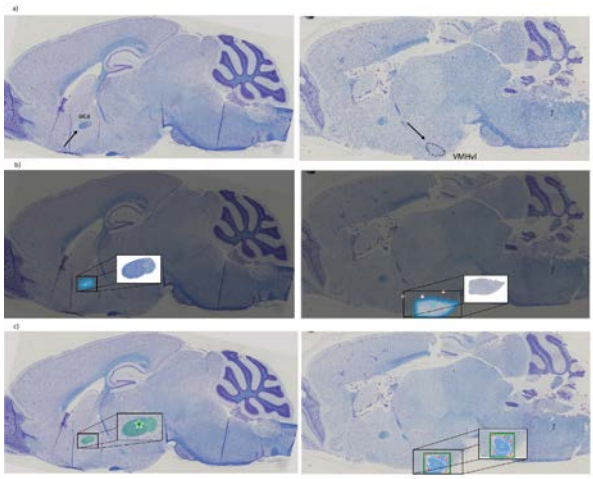


Figure 28: Segment Anything Model (SAM) output a) original image b) online demo c) with manual input seeds

The SAM performance for the two regions shows that for areas where there is a good contrast and they are easily differentiated (visually) from each other, the framework will perform well. Whereas, for regions where the difference is at the cellular level (e.i. VMHvl or InfC) its performance is not optimal.

4.7. Deployment

The proposed model was implemented, with a 7-level deep Attention U-Net, on several computers to test its performance. The following results were obtained and are shown below:

- CPU and HDD
  4800 sec/image (AMD RYZEN 7 3700U CPU @ 2.30 GHz, 24 GB RAM, 8 THREADS)
- CPU and SSD
  2760 sec/image (AMD RYZEN 7 3700U CPU @ 2.30 GHz, 24 GB RAM, 8 THREADS)

  1350 sec/image (INTEL CORE i7-10870H CPU @ 2.20 GHz, 16 GB RAM,16 THREADS)
- GPU and SSD
  300 sec/image (GPU TESLA K80 12 GB)

After the whole process, the final result is as many ROI files as analyzed regions. It should be emphasized that a result is not always obtained, either because the area is not recognized correctly or because the region is not present in mouse brain. These files will be mainly used by the Fiji/ImageJ software to perform the different neuroanatomical studies.

5. Discussion

The proposed model is presented as a general approach for automatic segmentation of different regions of the mouse brain using histological images. The work began with a review of the state of the art of murine models in neuroanatomy studies and automatic segmentation systems focused on mouse brains working with histological images. The first part of the review allowed us to understand the extreme need to achieve good accuracy when annotating brain regions. With the correct annotations, genes that help to understand the progression of developmental diseases such as ADHD can be identified. The task of annotating the mouse brain requires practice and expertise for its correct use in future studies. Reducing annotation time is therefore a necessity. In the second part of the review it was first found that, a large amount of information was found for magnetic resonance imaging, but not for histologic imaging because these require a higher resolution and their manipulation is complicated. This restriction means that its study requires the appropriate tools (e.g. memory disk, graphics cards) to work and develop automated systems.

This proposed approach will serve as a basis for future work in relation to accurate generation of mouse brain landmarks. The study did not use any advanced deep learning method, thus proving that the tools required for its development is no longer a constraint and opening the way for future research, taking this as a reference. Two main aspects found during the realization

of the model are highlighted. The first is the drastic improvement in using attention gates to focus the segmentation and improve the results. The second is the low false positive ratio value which reflects that our model always looks where it should look, thanks to the attention gate. In the results obtained by the proposed model, Attention U-Net 7 levels deep, the best value obtained is 99.14% of DSC for the total brain area. This region is the largest area but has certain restrictions in terms of its correct annotation which are easy to identify manually, but which our model also did. On the other hand, the regions with the best performance such as the fibers of the pons, are regions where their identification is not so evident due to their grouping at the cellular level and that are not always present in the brain. The result for this region was 70.47 in DSC.%.

In addition, the performance of the model was tested with the nnU-Net framework, which has been implemented in some segmentation challenges, reaching the first place. Comparing both performances, it can be said that the proposed model is at the level of this framework in terms of segmentation of the different regions of the mouse brain. It was evaluated in the areas where our model had the worst performance and the results obtained are similar to those of the nnU-Net.

The model was developed in a block form in which each of its parts can be substituted, replaced or improved. What makes it a versatile tool. The parts of the model presented are: dataset preparation, deep learning and image post-processing. The NeuroGeMM lab will be implemented as a means of extra annotation and will serve for inter-observer study when performing brain annotations to better test its performance.

5.1. Difficulties

When working with histologic images, they need to have a resolution that allows them to differentiate from shapes of the regions of interest and changes in tonality, to clusters of small cells and vessels. Therefore, having this high resolution will increase their size proportionally. This was one of the major limitations in making the proposed model, in addition to causing several times the saturation of normal memory discs.

The second limitation found is the correct annotation of brain regions. Therefore, several steps of treatment and revision of the dataset were required in order to continue with the deep learning training. This is also the reason why a multiple class approach was not chosen, because the manually annotation can cause overlapping between regions. Moving from a multiple class to a multiple label approach.

5.2. Future work

In order to continue working with the proposed method, the number of images can be increased. These can be of mice of different age, sex and with some pathology that affects the anatomy. Another field to be investigated is the implementation of multiple class approaches instead of binary approaches. But, before that, the dataset should be evaluated and corrected to avoid overlapping between the annotated regions.

6. Conclusions

The proposed model provides a starting point for the investigation of more accurate histological image segmentation systems. The model, apart from being easy to manage, does not require any additional software or training of the laboratory staff to use it. The system accepts as input a given type of images and converts them into landmarks of the regions of interest in the mouse brain. It takes as little as 5 minutes to correctly detemine 24 regions, which previously took an average of 1 hour to do the same task manually. The final results are ROI files of the analyzed regions, which will be used mainly by the Fiji/ImageJ software to perform the different neuroanatomical studies in the NeuroGeMM laboratory

Acknowledgments

This study is part of the Erasmus Mundus MAIA program as the final work of the master's degree. First, I would like to thank my family for their unconditional support. Also, I would like to thank my supervisors, Alain, Fabrice and Stephan, who knew how to guide me in the best way to complete this work. Finally, to the MAIA family for having made this master an unforgettable experience.

References

Barzekar, H., Ngu, H., Lin, H.H., Hejrati, M., Valdespino, S.R., Chu, S., Bingol, B., Hashemifar, S., Ghosh, S., 2023. Multiclass semantic segmentation to identify anatomical sub-regions of brain and measure neuronal health in parkinson's disease. arXiv preprint arXiv:2301.02925 .

Bland, J.M., Altman, D., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. The lancet 327, 307–310.

Bossert, L., Hagendorff, T., 2021. Animals and ai. the role of animals in ai research and application–an overview and ethical evaluation. Technology in Society 67, 101678.

Breschi, A., Gingeras, T., Guigó, R., 2017. Comparative transcriptomics in human and mouse. Nature Reviews Genetics 18. doi:10.1038/nrg.2017.19.

Bucher, M., Vu, T.H., Cord, M., Pérez, P., 2019. Zero-shot semantic segmentation. Advances in Neural Information Processing Systems 32.

Collins, S., Mikhaleva, A., Vrcelj, K., Vancollie, V., Wagner, C., Demeure, N., Whitley, H., Kannan, M., Balz, R., Anthony, L., Edwards, A., Moine, H., White, J., Adams, D., Reymond, A., Lelliott, C., Webber, C., Yalcin, B., 2019. Large-scale neuroanatomical study uncovers 198 gene associations in mouse brain morphogenesis. Nature Communications 10, 1234567890. doi:10.1038/s41467-019-11431-2.

Di Scandalea, M.L., Perone, C.S., Boudreau, M., Cohen-Adad, J., 2019. Deep active learning for axon-myelin segmentation on histology data. arXiv preprint arXiv:1907.05143 .

Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., et al., 2012. 3d slicer as an image computing platform for the quantitative imaging network. Magnetic resonance imaging 30, 1323–1341.

Flores, , Fullana, M., Soriano-Mas, C., Andero Galí, R., 2018. Lost in translation: how to upgrade fear memory research. Molecular Psychiatry 23. doi:10.1038/s41380-017-0006-0.

Hull, J.V., Dokovna, L.B., Jacokes, Z.J., Torgerson, C.M., Irimia, A., Van Horn, J.D., 2017. Resting-state functional connectivity in autism spectrum disorders: a review. Frontiers in psychiatry 7, 205.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods 18, 203–211.

Jahanifar, M., Tajeddin, N.Z., Koohbanani, N.A., Rajpoot, N.M., 2021. Robust interactive semantic segmentation of pathology images with minimal user input, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 674–683.

Kazdoba, T., Leach, P., Crawley, J., 2016. Behavioral phenotypes of genetic mouse models of autism. Genes, Brain and Behavior 15, 7–26.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al., 2023. Segment anything. arXiv preprint arXiv:2304.02643 .

Majdak, P., Ossyra, J.R., Ossyra, J.M., Cobert, A.J., Hofmann, G.C., Tse, S., Panozzo, B., Grogan, E.L., Sorokina, A., Rhodes, J.S., 2016. A new mouse model of adhd for medication development. Scientific reports 6, 1–18.

Mesejo, P., Ugolotti, R., Cagnoni, S., Di Cunto, F., Giacobini, M., 2012. Automatic segmentation of hippocampus in histological images of mouse brains using deformable models and random forest, in: 2012 25th IEEE International Symposium on Computer-Based Medical Systems (CBMS), IEEE. pp. 1–4.

Mokrzycki, W., M, S., 2012. New version of Canny edge detection algorithm. pp. 533–540.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 .

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, pp. 234–241. doi:10.1007/978-3-319-24574-4_28.

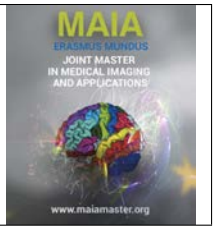
Rosset, A., Spadola, L., Ratib, O., 2004. Osirix: an open-source software for navigating in multidimensional dicom images. Journal of digital imaging 17, 205–216.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al., 2012. Fiji: an open-source platform for biological-image analysis. Nature methods 9, 676–682.

Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C., Gerig, G., 2006. User-guided 3d active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 31, 1116–1128.

Medical Imaging and Applications

Master Thesis, June 2023



Neurodegeneration identification in Parkinson's Disease with Deep Learning models using 3T quantitative MRI maps

Alejandro Cortina Uribe, David Meder

Danish Research Centre for Magnetic Resonance, Copenhagen, Denmark

Abstract

Parkinson's disease (PD) is a neurodegenerative syndrome with diverse motor and non-motor symptoms. While clinical assessment is the primary diagnostic method, magnetic resonance imaging (MRI) has gained importance in aiding PD diagnosis and treatment planning. While researchers have identified spatial patterns of neurodegeneration related to iron and neuromelanin (NM) that correlate with specific symptoms at 7T field strength, the applicability of these insights at 3T remains uncertain. Quantitative MRI (qMRI) maps are commonly used to model parameters that are robust across imaging sites and acquisition times. In PD, R2* and quantitative susceptibility mapping (QSM) images, highly sensitive to iron, are frequently employed. From a cohort study in our centre, we acquired 3T scans from which we can obtain different qMRI maps. Since the 3T protocol was not developed for PD imaging, performing frequentist statistics may not be suitable, and a DL-based analysis could provide better insights leveraging more powerful feature extraction and representation techniques.

Our study aims to investigate the ability of 3T qMRI maps to identify neurodegenerative changes in PD patients by training a well-performing DL pilot model using limited data and employing different learning techniques. We pursued two strategies: a) transfer learning-based binary classification using a 3D convolutional neural network (CNN) and application of explainable artificial intelligence (XAI) algorithms to interpret model predictions, and b) normative modeling, where we derived anomalies from reconstruction error maps and conducted binary classification based on the percentage of anomaly within specific regions of interest (ROIs). Although the first strategy did not yield a high-performing model, XAI proved invaluable in detecting issues such as overfitting and shortcut learning. In the second strategy, we performed group average statistics on reconstruction error maps and identified relevant subcortical nuclei in the MTsat, PD*, and R2* maps. By leveraging these ROIs, we quantified the error distribution among healthy controls and discovered anomalies that facilitated classification between PD patients and controls. The most discriminatory ROIs were the left globus pallidus interna in the MTsat map (AUROC: 0.84, G-mean: 0.82) and the left subthalamic nucleus (AUROC: 0.84, G-mean: 0.85). Our results highlight the challenges of binary classification with a small dataset and a 3D model architecture, even when employing diverse transfer learning strategies. However, the use of XAI to assess model predictions and identify signs of shortcut learning is crucial. Additionally, other learning techniques, such as unsupervised normative modeling, exhibit promising results, but necessitate careful selection of generative models, enlargement of the controls dataset to better capture its distribution, and rigorous validation of results.

Keywords: Parkinson's Disease, quantitative MRI, Deep Learning, normative modeling, classification, explainable AI

1. Introduction

1.1. Parkinson's Disease and Imaging

Parkinson's disease is a neurodegenerative syndrome that affects multiple motor and non-motor neural circuits. It involves two primary pathological processes: the loss of dopamine neurons and the accumulation of Lewy bodies. However, the order of occurrence of these processes is still unclear (Rizek et al., 2016). The loss of dopaminergic function leads to a decline in motor func-

tion and the emergence of clinical symptoms. Since there is no definitive test for confirming the diagnosis of PD, clinical diagnosis relies on assessing symptoms and patient history (DeMaagd and Philip, 2015). Neuroimaging studies, such as transcranial Doppler ultrasonography, PET, SPECT, and MRI, are performed to aid in the differential diagnosis and exclude other parkinsonian disorders (Rizek et al., 2016).

Structural changes resulting from neurodegeneration can be reflected in alterations in the local iron and neuromelanin (NM) content within the dopaminergic substantia nigra pars compacta (SNc) and the noradrenergic locus coeruleus (Madelung et al., 2022; Zucca et al., 2017) . Specifically, NM accumulates with age in the SNc within dopamine and noradrenaline neurons, but it depletes in PD patients due to the loss of these NM-containing neurons. On the other hand, iron also accumulates with age, but its deposition is excessive in PD (Biondetti et al., 2020; Zucca et al., 2017). These changes are strongly associated with motor impairment, such as the volume decrease of SNc in iron-sensitive quantitative susceptibility mapping (QSM) correlating with the severity of bradykinesia and rigidity, especially in patients with longer disease duration (Poston et al., 2020). Additionally, they are related to non-motor impairment, such as orthostatic changes in systolic blood pressure and apathy in locus coeruleus spatial neurodegeneration assessed by NM-sensitive MRI (Madelung et al., 2022).

Nevertheless, the relationship between these structural changes and the complex pathophysiology of PD is still not fully understood (Zucca et al., 2017). Magnetic Resonance Imaging (MRI) has become a valuable tool for researchers and clinicians to localize these changes, utilizing techniques such as NM-MRI (Trujillo et al., 2017) and iron-sensitive MRI (Biondetti et al., 2021). In recent years, high-resolution images obtained with ultra-high field scanners (7 teslas) have provided new insights into the topographic patterns of disease-related structural changes within these small nuclei ((Madelung et al., 2022). Furthermore, task-related functional MRI (fMRI) has revealed alterations in brain activation patterns related to the complex interactions of dopaminergic neurodegeneration in target nuclei (Meder et al., 2019).

However, the current MRI modalities targeting NM and iron have not yet provided robust diagnostic biomarkers for PD, mainly because they lack specificity to the melanin-iron complex or its metabolic processes during disease progression and onset. Additionally, research-only ultra-high field scanners are not widely available compared to the more commonly used 3 tesla MRI scanners, and it remains unclear whether MRI images acquired at this field strength can reveal similar or different patterns of PD-related changes.

Therefore, there is growing interest in emerging techniques such as quantitative MRI (qMRI) mapping, which aim to image tissue microstructure by modeling specific parameters (e.g., relaxation rates R1 or R2*), providing absolute measures and facilitating inter-site comparability across different time points (Tabelow et al., 2019; Weiskopf et al., 2013; Wenger et al., 2021). The most widely used quantitative maps in recent PD research are based on iron quantification within tissues, including T2* relaxometry (R2*) and quantitative susceptibility mapping (QSM) that utilize local susceptibility and phase information from gradient-echo or SWI sequences (Arribarat and Péran, 2020; Bae et al., 2021). In terms of NM imaging, these sequences exploit the property of melanin to reduce T1 relaxation time, while magnetization transfer imaging (MTw) is used to improve the contrast to NM, resulting in high-intensity signals in NM-rich areas (Bae et al., 2021; Madelung et al., 2022). Although quantitative maps derived from these sequences have not been extensively utilized, it is expected that R1 and magnetization transfer saturation maps contain information sensitive to NM.

1.2. Data analysis

To gain a better understanding of the aforementioned structural changes or functional patterns and draw interpretable conclusions, the field of neuroscience research has focused on conducting frequentist statistics on smaller cohorts. These cohorts are often limited by factors such as the availability of image modalities, subject and patient recruitment, and the complexity of disease progression.

More recently, deep learning (DL) has emerged as an alternative approach by addressing the problem of representation learning. DL aims to disentangle high-dimensional data into a lower-dimensional representation, enabling the identification of meaningful patterns and anomalies. In other words, DL attempts to learn abstract patterns that are relevant to the data.

Among the various learning problems that DL can assist with, classification tasks have been widely implemented. By training models to automatically extract features and perform "patient versus healthy control" classification for different brain diseases, we can develop end-to-end computer-aided diagnosis (CAD) systems that demonstrate exceptional predictive power compared to traditional machine learning models (see Section 2 State-of-the-art).

However, as we increase the complexity and flexibility of DL models, their interpretability and explainability diminish, contributing to the general skepticism among clinical researchers towards the "black box" nature of DL models. To address this concern, numerous explainability algorithms have been developed to gain insights into the learned features and decision-making processes of the models (Chaddad et al., 2023). Furthermore, the application of DL models in the medical domain is limited by data scarcity, which hampers their performance in generalization across different domains.

To mitigate this limitation, various training methodologies, such as transfer learning, unsupervised learning, and self-supervised learning, have been widely employed (Chen et al., 2019; Kim et al., 2022; Taleb et al., 2020).

Another valuable application of DL is the creation of normative models. In this framework, we move away from the assumption that clinical groups are easily distinguishable and homogeneous, aiming to better understand differences in relation to a reference model (Rutherford et al., 2022). Normative models have been utilized in various clinical scenarios, ranging from growth charting in pediatrics to mental disorders (Marquand et al., 2019). In the context of brain imaging, normative modeling has been employed to identify regions of the brain affected by disease or specific pathological patterns (see Section 2 State-of-the-art).

1.3. Project proposal

In this thesis project, we aim to investigate the relevance of qMRI maps acquired at 3 teslas in identifying structural changes in PD patients using a data-driven approach. We explore the possibility of training a high-performing DL pilot model with various learning techniques on limited data and examine the explanations for their performance. Our main general hypothesis is as follows:

- The qMRI maps (R1, R2*, PD*, and MTsat) obtained at 3 teslas are sensitive to neurodegeneration markers in PD, such as iron accumulation and NM loss, as well as potentially other structural changes. We will evaluate the classification performance of the proposed DL models and utilize explainability methods to identify relevant regions of interest.

We propose two exploratory strategies:

a) Unimodal binary classification with transfer learning: From a best performing model amongst different experiments we will initially obtain a predictive performance metric. Subsequently, by employing explainability methods, we will generate attribution heatmaps to localize the most important brain regions for the model's predictions. This approach may help us identify known nuclei affected by neurodegeneration, such as the SNc, as well as other regions of interest.

b) Normative modeling with unsupervised learning: In contrast to the first strategy, from PD patients we will first generate a reconstruction error map to identify disease anomalies and their spatial distribution. Then, by determining optimal thresholds that differentiate PD patients from controls, we will derive a final classification performance metric.

These two strategies will enable us to assess the potential of qMRI maps at 3 teslas in detecting structural changes related to PD. It is important to note that, despite the obtained qMRI maps were not particularly developed to be sensitive to PD neurodegenerative markers, we are optimistic that R2* maps are indeed sensitive to iron and MTsat and R1 maps might be sensitive to NM. This motivated our data-driven exploratory project oriented to investigate the sensitivity of these novel maps to identify structural changes related to PD, through DL methods that are able to capture and extract complex features and information from the images.

1.4. Abbreviations

PD, Parkinson's Disease. HC, healthy control. NM, neuromelanin. SNc, substantia nigra pars compacta. qMRI, quantitative magnetic resonance imaging. XAI, explainable artificial intelligence. ROI, region of interest.

2. State of the art

In our literature review, we did not find specific approaches that predicted PD or assessed PD neurodegeneration using qMRI maps and DL models. Currently, the research on DL-based PD classification has predominantly utilized other MRI sequences, brain imaging modalities such as SPECT and ECG, clinical and genetic data, or combinations of them.

When dealing with PD, the options for utilizing DL models are limited due to requirements of the dataset size. Thus, researchers often resort to using large public datasets like the multi-modal longitudinal Parkinson's Progression Markers Initiative (PPMI) (Marek et al., 2018) to train the models and validate them on smaller in-house datasets. In Chaki and Woźniak (2023), a systematic review highlighted that DL has been extensively used for neurodegenerative disorders in recent years. However, for PD, they found a majority of papers focusing on classification using non-brain imaging datasets such as speech and handwriting, with only a few studies using brain imaging data.

More recently, an increasing number of papers have been published using the PPMI study and other datasets to perform classification and explainability analyses. For instance, Camacho et al. (2023) gathered 13 different datasets comprising T1-weighted MRI scans (over 2000 participants). They employed a convolutional neural network (CNN) to classify PD and healthy control (HC) subjects using Jacobian maps derived from deformation fields of MNI spatial normalization, along with basic clinical parameters. They achieved an AUROC of 0.86 in their independent test set and generated saliency maps using the SmoothGrad (Smilkov et al., 2017) algorithm, which identified frontotemporal regions, the orbital-frontal cortex, and multiple deep gray matter structures as the most important.

In Shinde et al. (2019), an in-house dataset of 80 subjects of NM-sensitive MRI was used to classify PD

patients versus HC subjects and PD versus parkinsonian syndromes (APS) patients. They employed a 2D ResNet50 model trained with axial slices of the brainstem region. Their classification results were compared with two other ML-based models using radiomics and contrast-ratio features, and they obtained an AUROC of 0.906 on their test set, outperforming the ML approaches (AUROC of 0.54). To explain the DL model's decisions, they created class activation maps (CAM) from the weights and feature maps of the last convolutional layers and assessed contra-lateral activations in the SNc. They found a significantly larger mean activation in the left SNc compared to the right in PD patients.

In Huang et al. (2023), to address the limited interpretability of DL models, the authors defined disease classification (prodromal PD versus HC) as a graph representation task. They obtained relevant clinical interpretations by highlighting key nodes. They used diffusion tensor imaging (DTI) and structural MRI data from 194 subjects in the PPMI dataset to track fiber tracts and construct structural brain networks (SBNs). By employing a graph neural network, they achieved promising classification performance compared to other DL-based and ML-based models. Furthermore, through parametric decomposition and leveraging embedded GNN characteristics, they identified salient structural regions of interest (ROIs) that occurred most frequently among subjects, highlighting diverse cortical structures such as the precentral gyrus-L and the superior frontal gyrus-orbital.

The normative modeling framework has gained interest in recent years for its application in medical imaging tasks such as segmentation and classification. Additionally, it has been explored as a means to detect anomalies and identify lesions in brain MRI (Tschuchnig and Gadermayr, 2021). For instance, in the study by Baur et al. (2019), a novel deep autoencoding model with adversarial training was proposed for the detection and delineation of multiple sclerosis (MS) lesions based on reconstruction error maps. The authors trained a variational autoencoder (VAE) using 2D slices of FLAIR images from an in-house dataset of 83 healthy subjects, achieving the highest dice score coefficient (DSC) compared to other model architectures.

Similarly, Pinaya et al. (2021a) employed autoencoders to identify deviations from normal brains in Alzheimer's disease (AD) patients and identify associated critical regions. They trained a conditional autoencoder on a large cohort of healthy controls using subregional volume features extracted from over 11,000 structural MRI images from the UK Biobank. The performance of the model was validated on five additional datasets, where the mean squared error (MSE) between the reconstructed and inputted data served as a metric for brain deviation. This approach demonstrated high discriminative performance in distinguishing between healthy controls and AD patients. In a subsequent study, Pinaya et al. (2021b) developed a novel model based on VAEs and transformers to automatically detect various types of lesions and their delineations. By training their normative model on 15,000 FLAIR images from the UK Biobank, they achieved superior performance in lesion detection, specifically for white matter hyperintensities and tumors, outperforming similar autoencoder-based models in terms of DSC.

Lastly, in Muñoz-Ramírez et al. (2022), they identified subtle anomalies in *de novo* Parkinsonian patients by training spatial autoencoders with healthy controls DTI scans from the PPMI dataset. By utilizing 2-channel hemisphere axial slices derived from mean diffusivity (MD) and fractional anisotropy (FA) parameter maps, the authors generated joint reconstruction error maps for both the healthy control test set and the Parkinson's disease (PD) set. By evaluating the error maps per ROIs, they performed classification between controls and patients, achieving the highest geometric mean (G-mean) value for the macro regions of white matter and temporal lobe, as well as subcortical structures including the globus pallidus interna and thalamus.

With all the previously mentioned approaches, we want to highlight the diverse MRI sequences and DL models that have been used, as well as the efforts to explain the model's decisions and find disease-relevant ROIs. The latter aspect is particularly crucial when employing DL-based classification models, as the localization of spatial neurodegenerative patterns is essential in the current clinical diagnostic strategy. Therefore, explainability algorithms that provide attribution heatmaps at the pixel-level are necessary. In normative approaches, this region localization is inherently obtained through the reconstruction error map. Moreover, it is evident from these studies that either large cohort datasets or the extraction of 2D slices from MRI images are commonly employed to account for the size of the used dataset. Given the limitations of a small dataset in the present project, we adopt various learning techniques to evaluate the potential of qMRI maps in identifying PD neurodegeneration compared to more widely used MRI sequences while preserving the 3D nature of MRI scans and exploit inter-slice information to extract valuable information.

3. Material and methods

The general structure of this project comprised the following. Initially, we used an existing internal dataset part of the 7TPD project of the Danish Research Centre for Magnetic Resonance (Madelung et al., 2022), composed of 7T and 3T structural MRI of PD patients and HC subjects. From the 3T data we obtain a series of qMRI maps and, according to the requirements of the following steps, we pre-processed them (e.g. intensity rescaling, skull stripping, etc.). Then, we developed the proposed strategies of work: a) a series of ex-

periments systematically designed to compare different pre-training techniques and perform binary classification with DL models with the general goal of obtaining the best-performing pilot model, and later implement XAI algorithms to obtain attribution heatmaps; and b) a series of experiments designed to perform normative modeling of 3D neuroimaging data of a healthy population, thus creating an anomaly detector for PD patients. Each branch posed different challenges and limitations that will be addressed accordingly.

3.1. Dataset

We had access to a dataset of MRI scans acquired at 3 teslas on a Siemens Magnetom Prisma 3T scanner, comprising the multi-parameter mapping (MPM) protocol proposed by Weiskopf et al. (2013). The MPM protocol includes three multi-echo 3D fast low-angle shot (FLASH) scans: proton density (PDw), T1w, and magnetization transfer (MTw), a map of the B0 field (double gradient-echo FLASH acquisition) and a series of 3D EPI acquisitions of spin-echo (SE) and stimulated echo (STE) to map the RF transmit field B1. The dataset includes 72 subjects, out of which 49 have been diagnosed with PD and 23 are healthy controls (HC). In the PD group, there are 21 (42.85%) females and 28 males, with a mean age of 65± 10.75 years, and in the HC group, there are 8 (34.78%) females and 15 males, with a mean age of 67 ± 9.07 years.

We used the hMRI toolbox (Tabelow et al., 2019) that is based on SPM12 to obtain 1 mm high-resolution qMRI maps (Fig. 1):

- Longitudinal relaxation rate (R1 = 1/T1)
- Effective proton density (PD*)
- Magnetization transfer saturation (MTsat)
- Effective transverse relaxation rate (R2* = 1/T2*)

We used the multi-echo (TE = 2.34, 4.68, 7.02, 9.36, ..., 14.04 ms) FLASH scans: six MTw, eight PDw echoes, and eight T1w, to model their signal by the Ernst Equation (Ernst and Anderson, 2004), thus obtaining R1, PD*, and MTsat maps. The R2* map was derived through log-linear weighted least squares (WLS). To correct the quantitative data for transmit bias, the B1 transmit bias field was determined using consecutive pairs of SE/STE images corresponding to different flip angle nominal values, as well as the B0 field magnitude and phase images. Also, we corrected the RF sensitivity bias through the Unified Segmentation method, since no RF sensitivity map from the body and/or head coil was available. For further explanation of the methodology, please refer to Tabelow et al. (2019).

After obtaining the qMRI maps from all subjects, visual assessment was performed and two subjects (i.e. PD group, both males) were discarded due to data corruption problems.

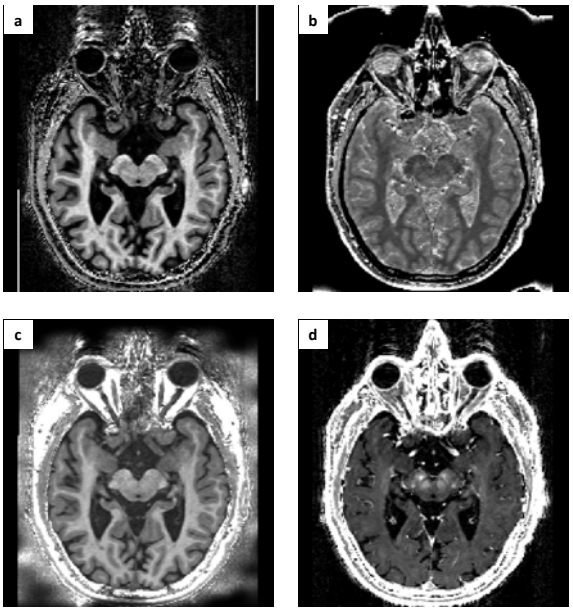


Figure 1: Quantitative MRI maps: a) MTsat, b) PD*, c) R1, d) R2*. Here displaying an axial slice at the SNc level after intensity scaling.

We preprocessed the obtained maps by first scaling the intensities to the recommended range per map: PD* = [50, 120] p.u., MTsat = [0, 2] p.u., R2* = [0, 70] s-1, R1 = [0, 1.4] s-1. After that, as required per each experiment level, we masked the volumes to obtain a region of interest accordingly. To obtain the skull-stripped volumes we utilized SynthStrip (Hoopes et al., 2022) and to obtain the brain parcellation we used SynthSeg (Billot et al., 2023), both tools available on FreeSurfer. From the brain parcellation, we had 33 labels from which we used the brainstem, left and right ventral diencephalon, left and right caudate, left and right thalamus, and left and right putamen, to create a binary mask of the brainstem and other nuclei of interest, which we will refer to from now on as the brainstem region.

For our comprehensive analysis and evaluation, we incorporated two labeled atlases: the previously mentioned SynthSeg atlas, which encompasses macroregions and selected subcortical parcellation regions, and the MNI PD25 atlas (Xiao et al., 2014), which specifically serves to MRI analysis and enables localization of pertinent PD regions (refer to the Appendix A.4 for a complete list of labels). The MNI PD25 atlas provides bilateral subcortical structure delineations, including the red nucleus (RN), substantia nigra compacta (SNc), subthalamic nucleus (STN), putamen, caudate, thalamus, and external and internal globus pallidus (GPi, GPe). To align each subject's R1 qMRI map with the PD25 T1 MPRAGE average atlas, we employed ANTs (Avants et al., 2011) for rigid, affine, and deformable spatial normalization. Cross-correlation served as the registration metric, and we used a multi-resolution framework to enhance the accuracy of the

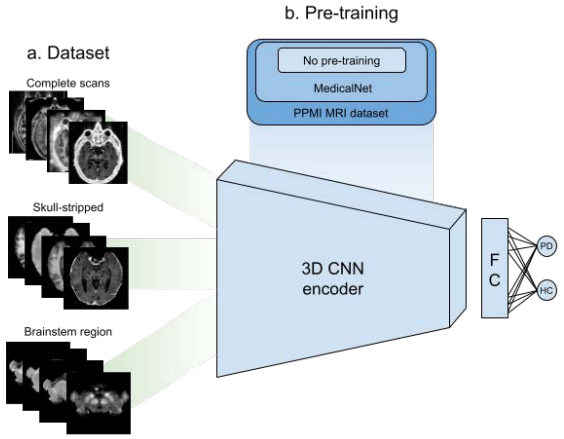


Figure 2: Binary classification strategy. a) Three different datasets per qMRI map type: complete scan, skull-stripped, and brainstem masked. b) Three different leves of model pre-training: no pre-training, using MedicalNet pre-trained model, and further pre-training using PPMI's MRI dataset of T1w images.

process.

3.2. Binary classification

We performed single-modality binary classification using transfer learning and a convolutional neural network (Fig. 2). We used the 3D Resnet (He et al., 2015) model architecture since it contains residual connections to tackle the vanishing gradient problem, and we chose the smallest version of that family to avoid over-parametrization.

Encoding full 3D scans and performing supervised binary classification requires a sufficient number of data samples to avoid overfitting the model or driving it to shortcut learning. Shortcut learning occurs when a model focuses on unintended easy-to-learn unrelated features, leading to a lack of generalization and unintuitive failures (Geirhos et al., 2020). To investigate this phenomenon, we conducted independent experiments where the model was trained with three distinct levels of region-of-interest (ROIs) (Fig. 2a). This strategy aims to limit the models to overfit to irrelevant spatial information or noise at each level.

Additionally, in the medical domain, transfer learning has emerged as a valuable technique to tackle limited data availability. This approach involves pre-training a model on a large-scale dataset, allowing it to extract general features, and subsequently fine-tuning it on a smaller dataset for the specific task at . We explored two levels of pre-training using medical datasets. Firstly, we leveraged the pre-trained models from MedicalNet (Chen et al., 2019), a framework trained on eight diverse medical image datasets (3DSeg-8), encompassing various imaging modalities such as MRI and computed tomography (CT). The authors have demonstrated notable performance improvements in segmentation and classification tasks using these models (Chen et al., 2019).

Subsequently, we extended the pre-training by incorporating MRI images from the PPMI dataset (Fig. 2b). The PPMI dataset (Marek et al., 2018) encompasses multimodal imaging data, including CT, fMRI, SPECT, PET, DTI, and MRI, collected at different time visits from two main cohorts: Parkinson's disease (PD) patients and healthy controls. For our purposes, we utilized the 3T 3D T1-weighted scans from the initial visit, resulting in a final dataset of 481 subjects (372 patients and 109 healthy controls). We utilized the MedicalNet pre-trained network and fine-tuned it using 60% of our PPMI dataset.

In this way, we have the same model architecture and three available sets of pre-trained weights (i.e. model parameters). We carried out transfer learning by replacing the pre-trained classification head with an adaptive max pool 3D layer and a single fully connected layer, with Xavier uniform parameter initialization. Because of this, for that group of parameters, we used an initial learning rate ten times larger than the group of parameters from the encoder.

Furthermore, we employed data augmentation techniques on the training set, a widely adopted approach to artificially expand the training set by applying various random transformations. The primary objective is for the model to encounter diverse variations of a single subject and learns robust features from them, for example, image orientation, rotations, or even changes in contrast. It is important to note that while traditional augmentation aims to create variations that align with the reality or the imaging technique's nature/domain, recent approaches have explored the opposite, generating synthetic data to enhance the model's robustness to various variations (Billot et al., 2023). Since our project focuses on assessing the predictive capabilities of qMRI maps, we employed simple affine transformations that would not modify the intensity content of the images as we were interested in preserving small or subtle contrast information.

For each experiment, we trained the model for a maximum of 150 epochs and implemented early stopping along with a reduce-on-plateau learning rate scheduler. To ensure optimal performance, we conducted a conservative hyperparameter tuning process, which involved evaluating different optimizers, loss functions, and initial learning rates. Considering the extensive number of experiments and the limitations of time and computational resources, we opted for a single train-validation stratified split of 80% and 20%, respectively.

3.2.1. Self-supervised learning

One of the most widely used strategies to face the limitations of data availability in the medical domain is to perform self-supervised learning (SSL). In SSL, opposite to traditional transfer learning strategies, the prelearnt features are derived from the same data through

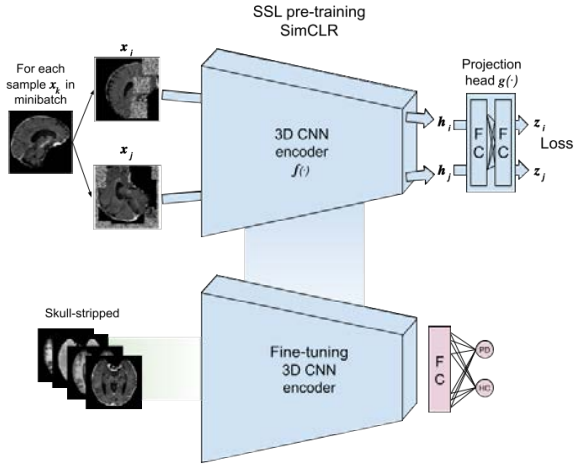


Figure 3: Self-supervised pre-training. Top: By using our own skull-stripped dataset, we train an encoder $f(\cdot)$ and a projection head $g(\cdot)$ using the SimCLR framework. Bottom: After pre-training, only the encoder is used for the downstream classification task using skull-stripped volumes.

proxy-task training. Subsequently, fine-tuning is performed in a supervised downstream task, reducing the need for a larger sample size and their annotations (Taleb et al., 2020).

The main goal of using this approach is to have the model learn an embedding space that is based on semantic similarity. In general, a spatial context proxy task is defined, such as predicting the relative position between image patches (Doersch et al., 2015), solving jigsaw puzzles (Noroozi and Favaro, 2017), or based on contrastive predictive coding (Hénaff et al., 2020; van den Oord et al., 2019). We chose to use the Simple Framework for Contrastive Learning of Representations (SimCLR) (Chen et al., 2020), since it has achieved state-of-the-art results in various computer vision tasks.

In the SimCLR framework we needed to follow two-steps. First, we created two different views from each image in the training dataset by using a heavy data augmentation composed of random flipping, affine transformations, and by masking regions of the image with noise (Fig. 3). With this, we were aiming for the model to encode information regarding the intensity distribution of different parts of the brain. Second, we trained a projection head in a contrastive manner by maximizing an agreement between differently augmented views of the same image while minimizing an agreement between views from different images. For this, we used the NT-Xent loss (Eq. 2) , which is a normalized temperature-scaled cross entropy loss that uses cosine similarity (Eq. 1).

$$\mathrm{sim}(z_i, z_j) = \frac{z_i^\top \cdot z_j}{\|z_i\| \cdot \|z_j\|} \tag{1}$$

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(z_i, z_k)/\tau)} \tag{2}$$

Where z_i, z_j are the embeddings output of the classification head coming from two augmented views of the same image, and τ is the temperature parameter.. We trained the model for 400 epochs using Adam optimizer and a learning rate of 0.001. After that, we used the SSL pre-trained network and fine-tune it for the unimodal binary classification task using only the skull-stripped volumes.

3.2.2. Explainability of Artificial Intelligence (XAI)

The primary objective of this project is to enhance the transparency of the model's predictions. To achieve this, we explored various XAI algorithms to gain insights into the model's behavior through attribution heatmaps and to gain a better understanding of disease-related spatial neurodegeneration. Typically, XAI methods are employed once a robust model with good performance and validated generalization is obtained, allowing for the assessment of any shortcut learning by visualizing relevant features.

To obtain feature importance attribution, we implemented two primary attribution algorithms: occlusion sensitivity (OS) and integrated gradients (IG), which evaluate the contribution of each input feature (voxel) to the model's output through image perturbation or manipulation.

OS is a method that involves masking or occluding parts of an input image to determine the contribution of each pixel to the output of a neural network. This method can help identifying the regions of an image that are most salient for a given classification task (Fig. 4a). With OS, we obtain an attribution heatmap at a pixel-level, meaning that we would know how much a region of specified size attributes to the model's final confidence score (Zeiler and Fergus, 2013). Because of this, the final resolution of the map depends solely on the sliding window size and stride, and furthermore, changing these parameters will influence directly the interpretation of the map regarding the relevance of the spatial information.

IG, on the other hand, computes the importance of each input feature for the neural network's output by integrating the gradient of the output with respect to the input along a straight path from a baseline input to the actual input (Fig. 4b). By integrating the gradient over this path, the method can capture the contribution of each feature to the final prediction (Sundararajan et al., 2017). In practice, the integral is efficiently approximated through summation, with the parameter m representing the number of steps between the baseline and the model's input.

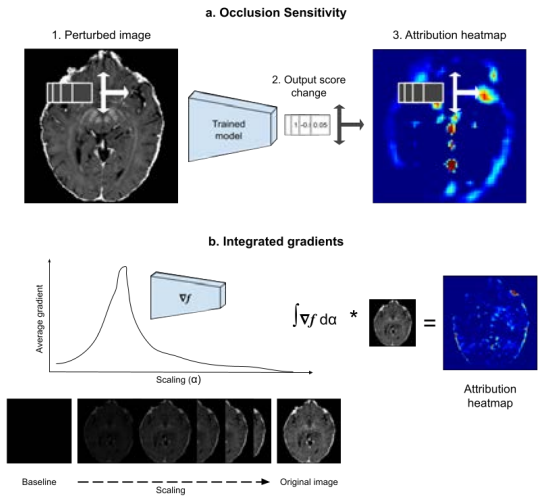


Figure 4: XAI. a. Occlusion sensitivity, by occluding parts of the image with a sliding window we measure how important that part is for the model, b. Integrated gradients, starting from an informationless baseline, the model gradients are computed and later integraded w.r.t. the scaling factor α.

For both algorithms, we utilized PyTorch's Captum implementation (Kokhlikyan et al., 2020). We determined a sliding window size of 8 voxels and a stride of 5 voxels as a suitable trade-off between granularity and computational cost for OS. Very small and overlapping patches significantly increase computation time when occluding a 3D volume. Regarding the IG algorithm, we used a zeros image as the baseline and approximated the integral using 200 steps.

For the best performing model, we obtained XAI maps for accurately predicted healthy control (HC) and Parkinson's disease (PD) subjects with the highest confidence scores, selecting a total of 8 subjects (4 HC and 4 PD) from the validation set. To identify the most significant between-group differences, we performed group average statistics per region of interest (ROI) using independent-samples t-tests and one-way ANOVA (F statistic) tests on the mean values derived from the normalized XAI maps.

3.3. Normative modelling

In our second line of investigation, we pursued an unsupervised learning approach using normative modeling to create a model of a healthy brain. Our goal was to identify variations from the norm in diseased brains. The basic concept involved constructing an identity model, where an original image served as input, and the model aimed to produce a reconstruction that closely resembled the original, thereby minimizing the reconstruction error (RE) between them. After the model is trained, when a pathological scan is provided as input, we expect to obtain a RE map indicating areas where the scan deviated from normality. This RE map functioned as an explanation heatmap for the model's predictions. Subsequently, by determining an optimal error threshold, we could evaluate the discriminative capabilities of different ROIs in distinguishing between diseased and control samples, enabling the computation of a performance metric.

As seen in section 2 State of the art, one of the most widely used architectures to perform normative modeling with brain imaging is the autoencoder (AE). In this simple structure, an image $x \in \mathbb{R}^{H\times W\times D}$ is fed through an encoder f_θ to obtain a latent space representation vector z, then a symmetrical decoder g_θ will then map z back to the reconstructed output $\hat{x} \in \mathbb{R}^{H\times W\times D}$. As concluded by Muñoz-Ramírez et al. (2022) and Baur et al. (2019), the dimensions of the latent space representation play a key role in the reconstruction error. Their experiments show that using a dense latent space $z \in \mathbb{R}^n$ performs significantly worse than having a 3D spatial latent space $z \in \mathbb{R}^{h\times w\times d}$, thus naming the autoencoder as spatial AE (sAE).

Although the AE can yield to high quality reconstructions, this type of model is not generative, meaning that as the model is allowed to create the latent space freely to output the best reconstruction, if we ever choose to create new synthetic images from a random latent embedding, we would obtain unrealistic noisy images. That is why variational autoencoders (VAE) were designed to mitigate this behaviour, as they map the original image to a latent space constraining it to follow a multivariate normal distribution, i.e. by encoding it into a mean μ and standard deviation σ latent variables. In this way, by sampling values from each variable we can obtain the latent space representation z.

To investigate how the latent space type and dimensions affects the reconstructions, we also implemented the vector-quantized VAE (VQ-VAE), a special type of VAE proposed by Oord et al. (2017). In it, the output of the encoder is mapped to the nearest point of a discrete latent space, so the latent embedding space is a codebook e of size K (i.e. vocabulary size) of vectors (i.e. words) with dimension D, $e \in \mathbb{R}^{K\times D}$. When training this framework (see details in Oord et al. (2017)), the codebook is learnt jointly with other model parameters. In order to obtain a final latent discrete representation, it would only be needed to replace each latent code by its index k from the codebook.

We employed fully convolutional 3D models to investigate the influence of depth and latent space size on the quality of reconstruction. In order to preserve the spatial information of the input data in the latent space, we opted for shallow encoders-decoders (Fig. 5a). Following the architecture of the VQ-VAE model proposed by Tudosiu et al. (Tudosiu et al., 2022), we implemented it using MONAI's Generative Models package (Cardoso et al., 2022). The VQ-VAE architecture incorporates residual units, where a selected number of residual blocks are placed after each convolutional layer. Each residual block consists of two consec-

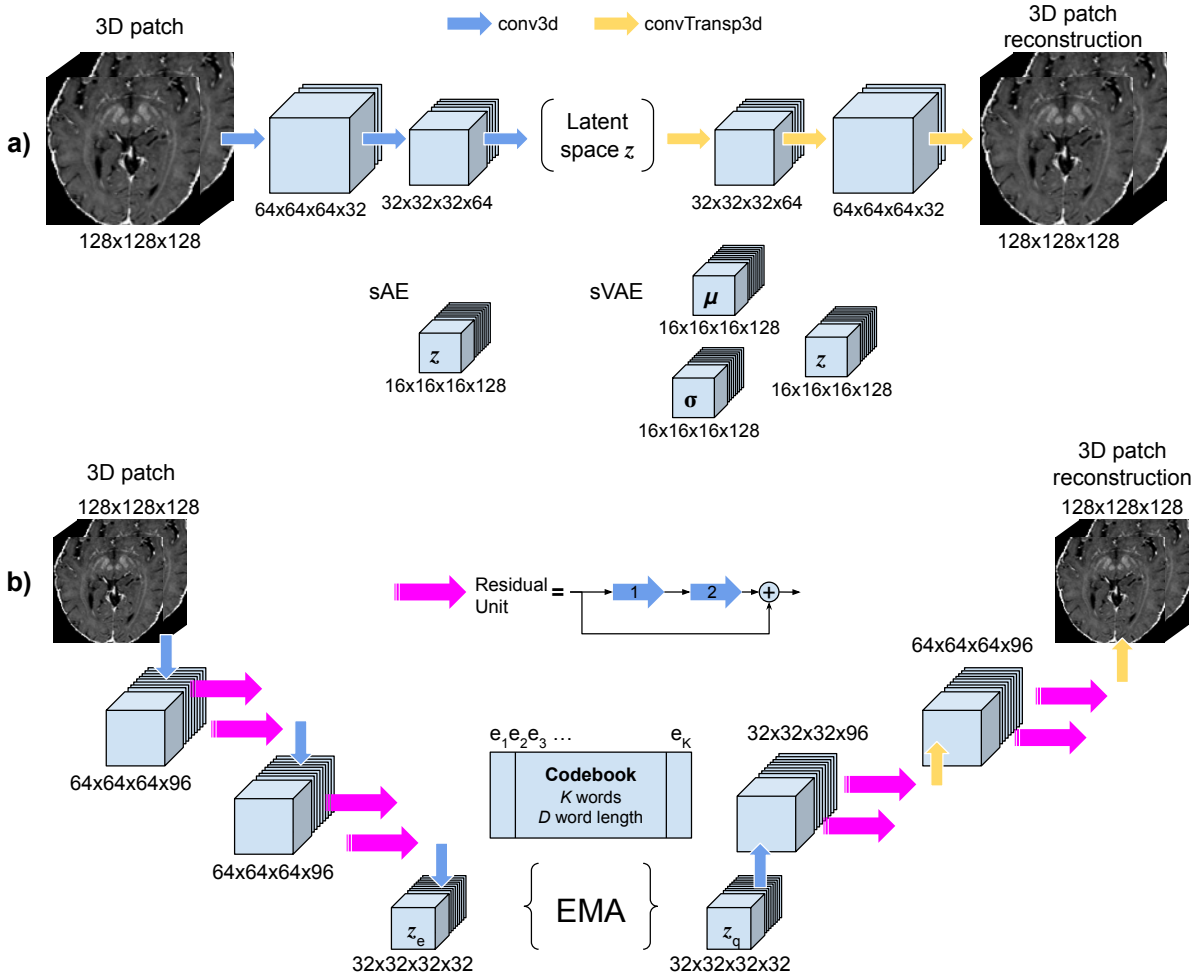


Figure 5: Normative modeling architectures. **a.** General architecture for the spatial autoencoder (sAE) and the spatial variational autoencoder (sVAE), only the latent space changes according to the type of model, z is the latent space embedding, μ and σ correspond to the normal variables. **b.** Vector quantized variational autoencoder (VQ-VAE) implemented architecture, the latent space embedding is vector quantized using a codebook of 32 words (K) of length 256 (D). The codebook is learnt along with the model's parameters using the algorithm exponential moving averages (EMA).

utive convolutional layers, with the output of the second layer being summed with the initial input. In our implementation, we used two convolutional layers in the encoder-decoder, with each layer followed by two residual blocks (Fig. 5b).

In order to augment our dataset, we adopted a patch-based approach for implementing our normative framework. This involved randomly cropping 3D patches from each volume, thereby introducing an additional parameter to consider. We chose a patch size of 128x128x128 to capture sufficient spatial information. To create the train and validation subsets, we split the HC set with a ratio of 70% for training and 30% for validation. For each training subject, we obtained nine patches from their respective volumes. During the inference phase, when testing a new image, we divided it into overlapping sub-volumes and fed each sub-volume to the model. The final reconstructed volume was then aggregated from all the sub-volumes, using a Hann window function to handle the overlapping regions and ensuring a smooth reconstruction.

For each of the model architectures, we used specific loss functions. In the simple sAE we used the L1 loss (Eq. 3). For the spatial VAE (sVAE), we used a loss function (Eq. 4) composed of the L1 norm as the reconstruction error and the Kullback-Leibler (KL) divergence to constraint the encoder to distribute all encodings around the center of the latent space (i.e. $\mu = 0$ and $\sigma = 1$). We weighted the KL term to favor the reconstruction term with a 0.9 ratio.

$$\mathcal{L} = \| x - \hat{x} \|_1 \tag{3}$$

$$\mathcal{L} = \lambda \parallel x - \hat{x} \parallel_1 + (1-\lambda)\left[-\frac{1}{2}\sum_{j=1}^{J}\left(1+\log(\sigma_j)^2-(\mu_j)^2-(\sigma_j)^2\right)\right] \quad (4)$$

Regarding the VQ-VAE training, we also used the L1 norm as reconstruction loss and the exponential moving averages (EMA) equation was used to learn the embedding space (i.e. learn the codebook parameters of the quantizer). With EMA, the embedding vectors e_i of the codebook are moved towards the encoder outputs $z_e(x)$. For the quantization loss details please refer to Oord et al. (2017). In the end, the final loss function comprised a sum of the reconstruction loss and the quantizer loss $\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{quant}$.

To assess the performance of our approach, we obtained reconstructed images and their corresponding RE maps from both the HC validation set and a sub-sample of PD subjects. The sub-sample of PD subjects was chosen to replicate the original imbalance ratio and match the sex distribution and average group age of the HC validation set. For the RE maps, we employed various measures of deviation, including the L1 norm, L2 norm, mean squared error (MSE), and the structural similarity index measure (SSIM) (Wang et al., 2004). The SSIM has been widely used in vision problems as it provides a better evaluation of perceptual image quality and structural similarity. Similar to our analysis in Section 3.2.2, we conducted group average statistics to examine between-group differences. Independent-samples t-tests and one-way ANOVA (F statistic) tests were performed on the mean and median RE values per ROI.

Finally, to evaluate the discriminant ability of each significant ROI, we established two thresholds. The first one, called abnormality threshold (a.t.), is set to detect abnormal voxels, serving for classification at voxel-level. We evaluated the a.t. as an extreme quantile value in the HC validation set error distribution. As noted by Muñoz-Ramírez et al. (2022), reconstruction errors can arise from various sources, such as data noise, loss of spatial information from the model, unaccounted variability in healthy controls, and actual anomalies caused by PD. Hence, selecting an extreme quantile (e.g., 98%) would classify only 2% of voxels in the control population as abnormal due to factors unrelated to PD. On the other hand, choosing a less restrictive quantile (e.g., 80%) would indicate that the model failed to accurately capture the distribution of controls, leading to the inclusion of genuine abnormalities within that threshold. Therefore, the a.t. can be considered a confidence threshold for the successful detection of abnormal voxels by the models.

Once the voxels in each ROI are thresholded based on the selected a.t., the proportion of anomalous voxels is determined, allowing the selection of the second threshold to evaluate the PD versus HC classification performance at the ROI level. Finally, receiver operating characteristic (ROC) curves are generated to assess the discriminating power of each ROI, and metrics such as the area under the curve (AUROC) and geometric mean (g-mean) are computed to quantify the classification performance.

4. Results

4.1. Binary classification

In this section, we present the results of our experiments in a sequential manner, allowing readers to follow the logical progression of our arguments throughout the experiments.

To evaluate the models' performance we utilized the area under the receiver operating characteristic curve (AUROC, or ROC-AUC) and the F1-score. We chose these metrics because they are more appropriate for imbalanced datasets compared to accuracy. The ROC curve plots the true positive rate ($TPR = TP/(TP + FN)$) or sensitivity/recall against the false positive rate ($FPR = FP/(FP + TN)$) or 1 - specificity at varying decision thresholds. With AUROC we measure the model's capability of distinguishing between classes and it ranges from 0 to 1. An AUROC of 0.5 indicates no separation capacity, above 0.5 indicates good separability, and below 0.5 indicates the model predicts the inverse class. The F1-score (Eq. 5) gives more weight to the positive class (i.e., PD) by not considering true negatives (TN). It is worth noting that in our sample, a scenario where the model incorrectly predicts all subjects as PD would yield a high F1-score (e.g., 0.78) due to the larger number of positive cases (Fig. 6).

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (5)$$

Confusion matrix		Predicted	
		PD	HC
Target	PD	9	0
	HC	5	0

= F1-score 0.783

Figure 6: Confusion matrix of the case where all subjects are predicted with the positive label PD and the high F1 score can be misleading.

Table 1 presents the classification results for all experiments per qMRI map, considering the three levels of pre-training and the three datasets used. The majority of experiments produced results similar to those depicted in Figure 6. However, some experiments demonstrated better performance in terms of high AUROC and F1-score, such as those involving the MTsat and R1 maps, utilizing complete scans and the PPMI pre-training level. For these experiments, we conducted

XAI analyses on selected subjects to gain further insights into the models' predictions and evaluate whether they had learned any shortcuts for the classification task (Fig. 7). The attribution heatmaps clearly reveal that the models learned to focus on information outside the brain, specifically in the neck and skull regions, respectively for the PD and HC examples.

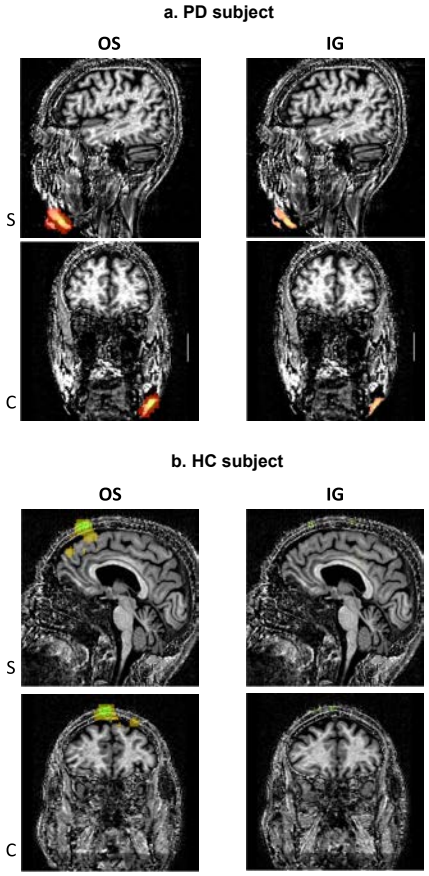


Figure 7: Occlusion sensitivity (OS) and integrated gradients (IG) XAI maps overlayed on two subjects' MTsat maps. Sagittal (S) and coronal (C) views were selected for better visualization. The heatmaps were obtained using the model trained with complete scans, and using the weights from the PPMI pre-train level. The heatmaps were thresholded to display positive attribution values and scaled for proper color intensity.

We then focused on the R2* map experiment using skull-stripped volumes and the PPMI pre-training level, which exhibited good performance in terms of AUROC and an improved F1 score. To gain insights into the model's decision patterns through the attribution heatmaps in a group analysis, we computed group average statistics for the subjects with the highest prediction scores. Specifically, we obtained the mean values for both OS and IG normalized attribution heatmaps for each ROI label in both the Synthseg parcellation and PD25 atlas. The ROIs that demonstrated p-values below our chosen confidence threshold ($\alpha < 0.05$) for both tests were considered the most significant (Fig. 8). The detailed results for all ROIs can be found in the Appendix A.12.

Among the most significant ROIs, in the IG attribution heatmaps we can identify some nuclei from the brainstem region for both atlases, highlighting that the pallidum (i.e. synthseg) and the globus pallidus interna and externa (i.e. PD25) are ROIs that significantly overlap and thus refer to the same region in the brain. On the other hand, in the OS attribution heatmaps, only the cerebral white matter macro-region and the lateral ventricles displayed a significant average difference between the groups.

Finally, in order to validate the performance of the R2* experiment, we performed 5-fold cross-validation(Table 2), clearly revealing that the model overfitted to that data split.

Table 3 presents the results of the SSL experiments. It is evident that across all maps, the models performed consistently, incorrectly predicting all subjects as PD and exhibiting low AUROC scores.

4.2. Normative modeling

Figure 9 presents an example of qualitative results for a R2* map, showcasing the reconstructed output for each model architecture and the different types of RE maps. Upon visual inspection, it is evident that the reconstructed outputs appear blurred for all models (Fig. 9a). In Figure 9b, we observe that lower values (close to 0) in the L1, L2, and MSE maps indicate fewer deviations from the normal brain, while in the case of SSIM, a higher value signifies greater similarity to the normal brain as it assesses structural similarity. For examples of MTsat, R1, and PD*, please refer to the Appendix A.13.

For each combination of qMRI map, model type, and RE type, we conducted group average statistics per ROI to assess the performance of the normative modeling approach. In the PD group, we anticipated observing an increase in mean or median error for L1, L2, and MSE, and a decrease in similarity according to SSIM. Figure 10 displays the ROIs from all experiments that exhibited significant p-values (i.e., $\alpha < 0.05$) for both statistical tests. We can see that each qMRI map showed at least one statistically significant ROI, with several subcortical nuclei being identified, including the right SNc in the R2* map. However, the R1 map only highlighted the left cerebral cortex as a relevant ROI. Furthermore, as expected, the error-based maps exhibited higher error values in the PD group, whereas the similarity-based map (Fig. 10a and d, right) unexpectedly showed higher values for the PD group. In cases where multiple RE map types and statistics (mean or median) yielded statistically significant results, we only report one per ROI.

For each of the significant ROIs identified by group average statistics, we evaluated the impact of the a.t. on the final performance evaluation and selected the ex-

(a) MTsat map

Dataset	Pre-training Level	AUROC	F1
Complete scans	None	0.811224	0.842105
	MedicalNet	0.913265	0.941176
	PPMI	**0.97449**	**0.941176**
Skull-stripped	None	0.678571	0.782609
	MedicalNet	0.69898	0.782609
	PPMI	0.739796	0.782609
Brainstem Region	None	0.678571	0.782609
	MedicalNet	0.709184	0.782609
	PPMI	0.739796	0.782609

(b) PD* map

Dataset	Pre-training Level	AUROC	F1
Complete scans	None	0.770408	0.782609
	MedicalNet	0.785714	0.8
	PPMI	0.714286	0.666667
Skull-stripped	None	0.739796	0.782609
	MedicalNet	0.760204	0.782609
	PPMI	**0.811224**	**0.782609**
Brainstem Region	None	0.709184	0.782609
	MedicalNet	0.668367	0.782609
	PPMI	0.655612	0.782609

(c) R1 map

Dataset	Pre-training Level	AUROC	F1
Complete scans	None	0.94898	0.888889
	MedicalNet	0.938776	0.888889
	PPMI	**0.933673**	**0.947368**
Skull-stripped	None	0.872449	0.782609
	MedicalNet	0.770408	0.782609
	PPMI	0.811224	0.782609
Brainstem Region	None	0.637755	0.782609
	MedicalNet	0.668367	0.782609
	PPMI	0.596939	0.782609

(d) R2*

Dataset	Pre-training Level	AUROC	F1
Complete scans	None	0.668367	0.782609
	MedicalNet	0.760204	0.782609
	PPMI	0.663265	0.727273
Skull-stripped	None	0.770408	0.782609
	MedicalNet	0.831633	0.782609
	PPMI	**0.94898**	**0.888889**
Brainstem Region	None	0.80102	0.782609
	MedicalNet	0.831633	0.782609
	PPMI	0.811225	0.782609

Table 1: Binary classification results for the validation set, per qMRI map. The best experiment's results per qMRI map are shown in bold.

Fold	AUROC	F1
1	0.918367	0.833333
2	0.595939	0.761905
3	0.729592	0.782609
4	0.529592	0.782609
5	0.69898	0.782609

Table 2: 5-fold cross-validation results for the R2* map, using skull-stripped volumes and PPMI level of pre-training.

Map type	AUROC	F1
MTsat	0.760204	0.782609
PD*	0.719388	0.782609
R1	0.760204	0.782609
R2*	0.760204	0.782609

Table 3: SSL pre-training classification results for the validation set, per qMRI map.

treme quantile that yielded the best result. For that selected a.t. we plotted the ROC curve and the G-mean (Eq. 6) and associated abnormality percentage (i.e. the second threshold that determines the optimal ROI-level classification) (Fig. 11). The highest classification results was achieved by the left globus pallidus interna (GPi) in the MTsat map (AUROC = 0.84, G-mean = 0.82) and the left subthalamic nucleus (STN) in the PD* map (AUROC = 0.84, G-mean = 0.85).

$$G - Mean = \sqrt{TPR * (1 - FPR)} \tag{6}$$

5. Discussion

We did not obtain satisfactory results for the binary classification strategy. However, by employing XAI techniques and conducting proper model validation, we gained valuable insights during the results analysis. Upon examining the XAI attribution heatmaps (Fig. 7), we might infer that the shortcut learning was due to structural information in the form of confounds (e.g. anatomical head variations that only one sample group showed), but because the qMRI maps showed very disrupted patterns outside the brain, we believe the model focused on learning noise. These findings highlight the importance of XAI in validating deep learning

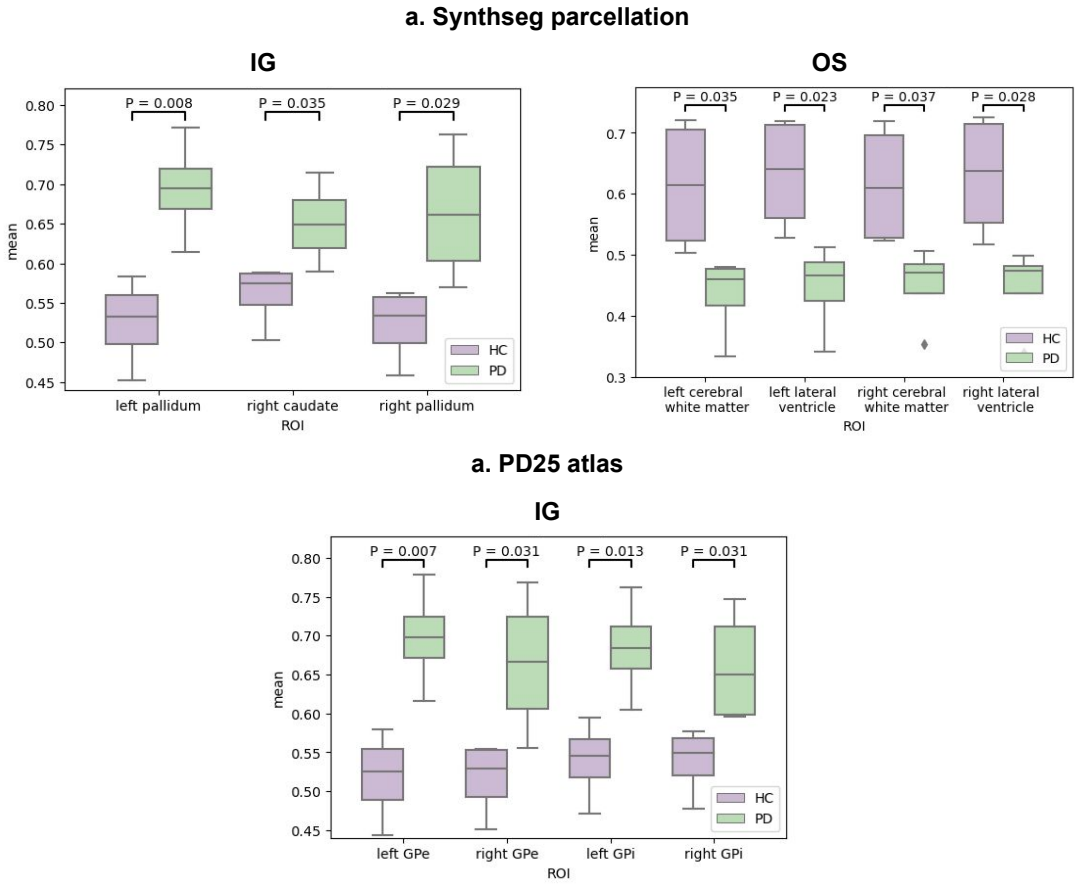


Figure 8: Statistically significant average difference at ROI level for the R2* experiment (PPMI pre-training level, using skull-stripped volumes). **a.** Integrated gradients (IG) and occlusion sensitivity (OS) results for Sythseg ROIs, **b.** IG results for PD25 atlas ROIs, OS had no significant results. Above each pair, the p-value associated to the ANOVA test is displayed, the p-value associated to the t-test was always 0.0285. Each group sampled contained the 4 subjects with higher predicted score. The XAI maps were normalized from 0 to 1 before computing the group statistics.

models' performance. Nonetheless, interpreting XAI attribution maps can be challenging, especially when higher attributions are found within the brain. It is important to note that inferring novel disease-related neurodegeneration without prior research would be difficult without specific hypotheses, as we have for PD and the SNc and LC ROIs. Nevertheless, by carefully examining the XAI maps, we confirmed that masking the volumes effectively eliminated regions where the model exhibited shortcut learning, thus reinforcing the need to gain a deeper understanding of our data to interpret the model's decision process.

In our analysis of group average statistics for the best R2* experiment (using skull-stripped scans and PPMI pre-training level), although it was latter shown with 5-fold CV that the model overfitted to that data split, we wanted to better understand the model decision and perhaps reveal similarly any shortcut learning evidence. However, adding the previously stated considerations, it is particularly difficult to infer explanations, mainly because there are other constraints to deal with when interpreting XAI attribution maps. For instance, ablation-based algorithms like OS, where some features are dropped and the change in predictions is noted, lead to unrealistic inputs and potentially misleading interpretations when features interact when changing the size of the occluded region (Sundararajan et al., 2017). This might explain the inclusion of the right and left lateral ventricles as relevant ROIs in the OS heatmaps (Fig. 8). Additionally, gradient-based XAI algorithms like Integrated Gradients (IG) can be easily manipulated by applying imperceptible perturbations to the input, making it difficult to interpret the resulting map as a reliable explanation or to use it for assessing our general hypothesis (Dombrowski et al., 2019).

We found that our pre-training strategies to address data scarcity did not yield satisfactory results for our problem. While in some experiments, such as PD* skull-stripped volumes (Table 1b) showed an increase in AUROC accordingly to the pre-training level, there was no clear pattern indicating consistent improvement across all qMRI maps and experimental settings. We might attribute this failure to two things: domain shift and data's high dimensionality. Although transfer learn-

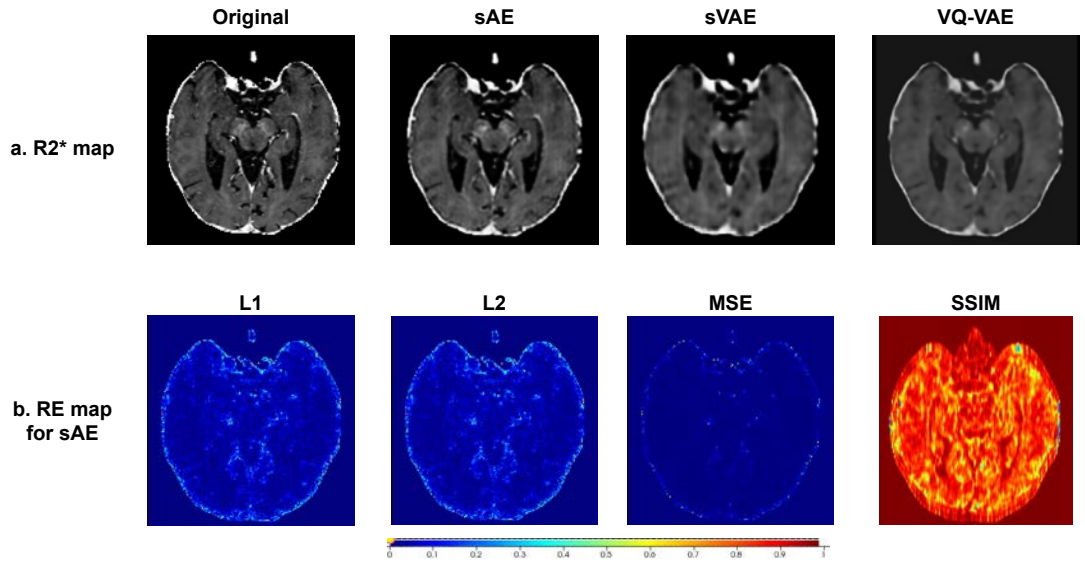


Figure 9: Qualitative results example for R2*. a. Reconstructed output for spatial autoencoder (sAE), spatial variational autoencoder (sVAE), and vector-quantized variational autoencoder (VQ-VAE). b. L1, L2, MSE, and SSSIM reconstruction error (RE) maps for sAE model.

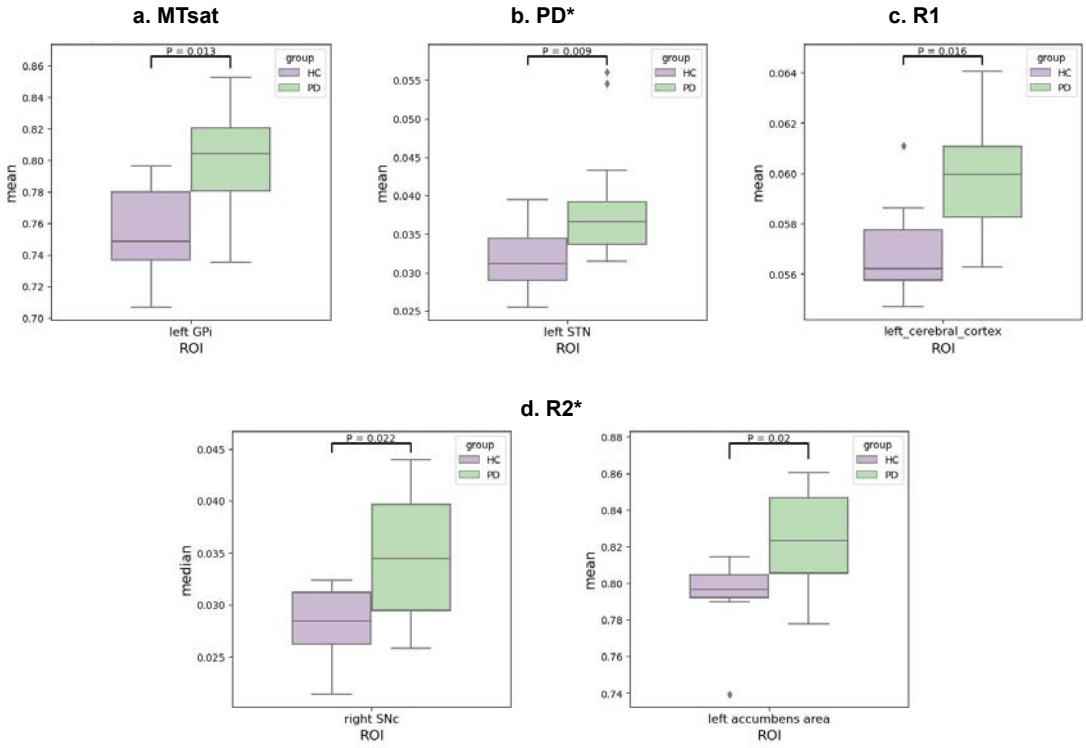


Figure 10: Statistically significant difference at ROI level for different normative modeling experiments. **a.** MTsat map, mean group differences using spatial autoencoder (sAE) and the SSIM RE map. **b.** PD* map, mean group differences using vector-quantized variational autoencoder (VQ-VAE) and the L1 RE map. **c.** R1 map, mean group differences using spatial variational autoencoder (sVAE) and the L1 RE map. **d.** R2* map, median group differences for sAE and L1 RE map (left), and mean group differences for sAE and SSIM RE map (right). Abbreviations: GPi - globus pallidus interna, STN - subthalamic nucleus, SNc - substantia nigra.

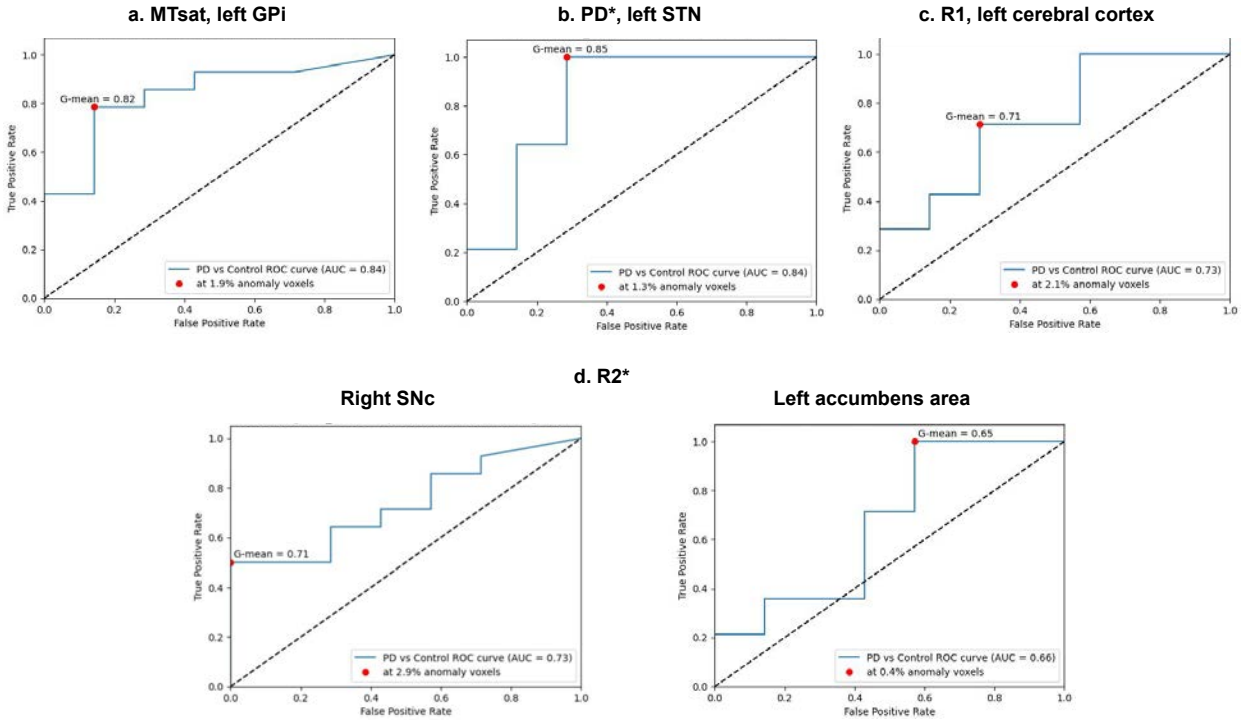


Figure 11: Classification results for normative modeling. Shown are the statistically significant ROIs and their associated ROC curve after selecting the appropriate a.t. In each ROC curve the AUROC is displayed, as well as the G-mean and corresponding abnormality percentage. **a.** Left GPi for MTsat map, at 99 quantile. **b.** Left STN for PD* map, at 98 quantile. **c.** Left cerebral cortex for R1 map, at 98 quantile. **d.** For R2* map, right SNc at 99 quantile (left) , and left accumbens area at 98 quantile (right).

ing in the medical domain has shown promising results for certain problems (Chen et al., 2019), domain adaptation is still an evolving field. In our case, the two pre-training stages were performed using imaging modalities different to the qMRI maps, and this domain shift between datasets prevented a proper transfer of learning. Additionally, we still require large amounts of data and our 3D strategy might not have been the most suitable for our dataset size. For our classification task, the limited number of samples and the high-dimensional nature of imaging data, combined with a single label per image, may have restricted the model's ability to make sense of the data as a whole. Furthermore, the poor results in our SSL experiments may be attributed to the choice of transformations for the augmented views, which failed to help the model learn the relevant features for the downstream task. Moreover, our explorative experiments were scarce and a patch-based framework could be explored to increase the dataset size, or even consider trying different proxy tasks, such as the jigsaw puzzle.

To determine which qMRI map was most suitable for our classification task, the only slight indication was of the brainstem region experiments, and that of the R2* map that showed higher AUROC scores compared to the other maps (Fig. 1d). However, since the model still predicted all subjects as PD (i.e., F1 score of 0.7826), we cannot definitively state that the R2* map contains better or more discriminatory information.

Regarding our normative modeling approach, we believe it was a better strategy for our problem for two main reasons. First, assessing a neurodegenerative disease such as PD as a continuum or a degree of deviation of normality better suits its diagnostic framework. Second, by generating RE maps we obtain explanation maps that allowed us to identify spatial patterns of anomalies, which was a primary goal of our project. Although the choice of error or similarity metrics to generate the RE map is crucial, it significantly reduces the limitations inherent in XAI algorithms, as we can directly interpret the error as a measure of abnormality.

Nevertheless, we still faced some limitations. The data scarcity problem still arised for this approach despite implementing a patch-based strategy to increase our dataset size. Our HC set was smaller than our PD set leading to a disproportionately smaller HC dataset, unlike most existing normative models (see section 2) that were trained with thousands of images. This directly affected our implementation when it came to choose the a.t., as it relates to the degree of confidence for the model to accurately learn the HC distribution. We can reasonably expect that the models failed to capture the full variability of healthy controls and likely overfit to patches from only 14 subjects. To properly validate our approach, it would have been necessary to sample the HC set and create different training-validation subsets.

Unfortunately, due to time constraints, we could not perform this validation.

Regarding our reconstruction outputs (Fig. 9), it is evident that our model architectures struggled to accurately reconstruct high-frequency features, resulting in blurred images. Despite our efforts to preserve spatial information by tuning the size of the latent space (z) and employing shallow fully convolutional networks, the sharpness of the reconstructions was limited by the chosen loss function. The use of L1 as a loss function inevitably drove the model to learn that a blurred image minimizes the error quickly. Furthermore, we attribute the degree of blurring to the constraints imposed by the models on the latent space distribution, whether it be following a multivariate normal distribution (sVAE) or being discrete (VQ-VAE), compared to the less restrictive sAE.

In our analysis of group average statistics, we found that when using L1, L2, or MSE, the error difference between the PD and control groups was higher than that within the control group, supporting the argument that the PD group deviates from the controls. However, we obtained contradictory results for the SSIM maps, where the similarity value should have been higher for the HC group compared to the PD group. This discrepancy may be attributed to the structural component of the SSIM, which is highly sensitive to edge and contour information—factors strongly affected by the blurry nature of the reconstructed outputs (Renieblas et al., 2017).

Finally, in the classification results of the normative modeling, although the left SNc in the R2* map showed significant group difference, it was not sufficiently discriminant to separate PD and HC in our test set, compared to other ROIs that showed better performance. This could be attributed to two reasons. First, we rely on the high contrast at voxel level in the R2* maps to assess iron deposition in PD, and since our models reconstructed blurred images, that high-frequency information was lost. Second, as mentioned before, the single-split training set may have not included enough controls to model the normal distribution of iron at the SNc, as well as it could have overfitted to a set of relevant examples.

6. Conclusions

In this thesis project, our goal was to explore the potential of deep learning (DL) models in uncovering novel insights into Parkinson's Disease (PD) neurodegeneration. To achieve this, we employed explainable artificial intelligence (XAI) algorithms to enhance the transparency of complex model decisions and identify relevant regions of interest. We initially pursued the traditional binary classification strategy, but encountered challenges in obtaining satisfactory results. However, this approach provided valuable insights, including the identification of shortcut learning, model validation and overfitting assessment, and the understanding of transfer learning capabilities and limitations.

In our second strategy, normative modeling, we achieved better-suited models for studying the disease and obtained intrinsic explainable reconstruction error maps that led to more interpretable conclusions. However, the results were modest due to the limitations of our generative models to adequately reconstruct important high-frequency information, and the lack of proper validation for the model's performance restrained us from making more profound interpretations.

Our intention was to leverage the high spatial resolution of MRI scans by employing 3D models. However, the small number of samples in our dataset suggests that implementing a 2D or 2.5D model would have been more appropriate. We also aimed to utilize pre-trained models and publicly available datasets, but we faced domain shift limitations to effectively transfer the learned knowledge to our specific classification task.

As future work, it would be beneficial to explore multi-modal strategies, such as combining the four quantitative MRI (qMRI) maps at 3T or integrating additional imaging data, such as 7T NM- and iron-sensitive images, as well as clinical data like PD scale ratings. However, careful consideration must be given to address the challenge of the curse of dimensionality and ensure proper interpretation and explanation of the model's decisions.

Moreover, it would be particularly interesting and clinically relevant to investigate multi-label or multi-class classification approaches, as PD encompasses a continuum of multiple motor and non-motor symptoms.

Acknowledgments

I would like to express my gratitude to Hartwig Siebner for providing me with the opportunity to carry out my thesis internship at the Danish Research Centre for Magnetic Resonance (DRCMR) and for his unwavering support in promoting transparency in research and fostering a passion for MRI. I am deeply thankful to David Meder for his valuable insights and expertise in elucidating the various pathological processes associated with Parkinson's Disease, which greatly contributed to the relevance of my project. I am indebted to the entire DRCMR staff for their generous assistance in answering my inquiries regarding MRI, data processing, and Parkinson's Disease, as their guidance was instrumental in the successful completion of this research endeavor. Special thanks go to José Bernal for his invaluable assistance in visualizing a more effective deep learning approach and to Vladyslav Zalevskyi for his unwavering availability, patient guidance, and for sharing this enriching experience with me. Lastly, I extend my appreciation to all the researchers who have made their research, code, and methodologies openly available, as

well as to the PPMI group for graciously granting me access to their invaluable imaging data.

References

Arribarat, G., Péran, P., 2020. Quantitative MRI markers in parkinson's disease and parkinsonian syndromes. Current Opinion in Neurology 33, 222–229. doi:10.1097/wco.0000000000000796.

Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C., 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. NeuroImage 54, 2033–2044. URL: https://doi.org/10.1016/j.neuroimage.2010.09.025, doi:10.1016/j.neuroimage.2010.09.025.

Bae, Y.J., Kim, J.M., Sohn, C.H., Choi, J.H., Choi, B.S., Song, Y.S., Nam, Y., Cho, S.J., Jeon, B., Kim, J.H., 2021. Imaging the substantia nigra in parkinson disease and other parkinsonian syndromes. Radiology 300, 260–278. doi:10.1148/radiol.2021203341. pMID: 34100679.

Baur, C., Wiestler, B., Albarqouni, S., Navab, N., 2019. Deep autoencoding models for unsupervised anomaly segmentation in brain MR images, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, pp. 161–169. URL: https://doi.org/10.1007/978-3-030-11723-8_16, doi:10.1007/978-3-030-11723-8_16.

Billot, B., Greve, D.N., Puonti, O., Thielscher, A., Van Leemput, K., Fischl, B., Dalca, A.V., Iglesias, J.E., 2023. Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining. Medical Image Analysis 86, 102789. URL: https://www.sciencedirect.com/science/article/pii/S1361841523000506, doi:https://doi.org/10.1016/j.media.2023.102789.

Biondetti, E., Gaurav, R., Yahia-Cherif, L., Mangone, G., Pyatigorskaya, N., Valabrègue, R., Ewenczyk, C., Hutchison, M., François, C., Arnulf, I., Corvol, J.C., Vidailhet, M., Lehéricy, S., 2020. Spatiotemporal changes in substantia nigra neuromelanin content in Parkinson's disease. Brain 143, 2757–2770. URL: https://academic.oup.com/brain/article/143/9/2757/5898381, doi:10.1093/brain/awaa216.

Biondetti, E., Santin, M.D., Valabrègue, R., Mangone, G., Gaurav, R., Pyatigorskaya, N., Hutchison, M., Yahia-Cherif, L., Villain, N., Habert, M.O., Arnulf, I., Leu-Semenescu, S., Dodet, P., Vila, M., Corvol, J.C., Vidailhet, M., Lehéricy, S., 2021. The spatiotemporal changes in dopamine, neuromelanin and iron characterizing Parkinson's disease. Brain 144, 3114–3125. URL: https://academic.oup.com/brain/article/144/10/3114/6274641, doi:10.1093/brain/awab191.

Camacho, M., Wilms, M., Mouches, P., Almgren, H., Souza, R., Camicioli, R., Ismail, Z., Monchi, O., Forkert, N.D., 2023. Explainable classification of parkinson's disease using deep learning trained on a large multi-center database of t1-weighted mri datasets. NeuroImage: Clinical 38, 103405. URL: https://www.sciencedirect.com/science/article/pii/S2213158223000943, doi:https://doi.org/10.1016/j.nicl.2023.103405.

Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., He, Y., Xu, Z., Hatamizadeh, A., Myronenko, A., Zhu, W., Liu, Y., Zheng, M., Tang, Y., Yang, I., Zephyr, M., Hashemian, B., Alle, S., Darestani, M.Z., Budd, C., Modat, M., Vercauteren, T., Wang, G., Li, Y., Hu, Y., Fu, Y., Gorman, B., Johnson, H., Genereaux, B., Erdal, B.S., Gupta, V., Diaz-Pinto, A., Dourson, A., Maier-Hein, L., Jaeger, P.F., Baumgartner, M., Kalpathy-Cramer, J., Flores, M., Kirby, J., Cooper, L.A.D., Roth, H.R., Xu, D., Bericat, D., Floca, R., Zhou, S.K., Shuaib, H., Farahani, K., Maier-Hein, K.H., Aylward, S., Dogra, P., Ourselin, S., Feng, A., 2022. Monai: An open-source framework for deep learning in healthcare.

Chaddad, A., Peng, J., Xu, J., Bouridane, A., 2023. Survey of explainable AI techniques in healthcare. Sensors 23, 634. URL: https://doi.org/10.3390/s23020634, doi:10.3390/s23020634.

Chaki, J., Woźniak, M., 2023. Deep learning for neurodegenerative disorder (2016 to 2022): A systematic review. Biomedical Signal Processing and Control 80, 104223. URL: https://linkinghub.elsevier.com/retrieve/pii/S1746809422006772, doi:10.1016/j.bspc.2022.104223.

Chen, S., Ma, K., Zheng, Y., 2019. Med3d: Transfer learning for 3d medical image analysis. doi:https://doi.org/10.48550/arXiv.1904.00625.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. doi:https://doi.org/10.48550/arXiv.2002.05709.

DeMaagd, G., Philip, A., 2015. Parkinson's disease and its management: Part 1: Disease entity, risk factors, pathophysiology, clinical presentation, and diagnosis. P T 40, 504–532.

Doersch, C., Gupta, A., Efros, A.A., 2015. Unsupervised visual representation learning by context prediction, in: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1422–1430. doi:10.1109/ICCV.2015.167.

Dombrowski, A.K., Alber, M., Anders, C.J., Ackermann, M., Müller, K.R., Kessel, P., 2019. Explanations can be manipulated and geometry is to blame. URL: http://arxiv.org/abs/1906.07983. arXiv:1906.07983 [cs, stat].

Ernst, R.R., Anderson, W.A., 2004. Application of Fourier Transform Spectroscopy to Magnetic Resonance. Review of Scientific Instruments 37, 93–102. URL: https://doi.org/10.1063/1.1719961, doi:10.1063/1.1719961.

Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A., 2020. Shortcut learning in deep neural networks. Nature Machine Intelligence 2, 665–673. URL: https://doi.org/10.1038/s42256-020-00257-z, doi:10.1038/s42256-020-00257-z.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. doi:https://doi.org/10.48550/arXiv.1512.03385.

Hoopes, A., Mora, J.S., Dalca, A.V., Fischl, B., Hoffmann, M., 2022. Synthstrip: skull-stripping for any brain image. NeuroImage 260, 119474. URL: https://www.sciencedirect.com/science/article/pii/S1053811922005900, doi:https://doi.org/10.1016/j.neuroimage.2022.119474.

Huang, L., Ye, X., Yang, M., Pan, L., hua Zheng, S., 2023. MNC-net: Multi-task graph structure learning based on node clustering for early parkinson's disease diagnosis. Computers in Biology and Medicine 152, 106308. URL: https://doi.org/10.1016/j.compbiomed.2022.106308, doi:10.1016/j.compbiomed.2022.106308.

Hénaff, O.J., Srinivas, A., Fauw, J.D., Razavi, A., Doersch, C., Eslami, S.M.A., van den Oord, A., 2020. Data-efficient image recognition with contrastive predictive coding. doi:https://doi.org/10.48550/arXiv.1905.09272.

Kim, H.E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M.E., Ganslandt, T., 2022. Transfer learning for medical image classification: a literature review. BMC Medical Imaging 22. URL: https://doi.org/10.1186/s12880-022-00793-7, doi:10.1186/s12880-022-00793-7.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., Reblitz-Richardson, O., 2020. Captum: A unified and generic model interpretability library for pytorch.

Madelung, C.F., Meder, D., Fuglsang, S.A., Marques, M.M., Boer, V.O., Madsen, K.H., Petersen, E.T., Hejl, A., Løkkegaard, A., Siebner, H.R., 2022. Locus Coeruleus Shows a Spatial Pattern of Structural Disintegration in Parkinson's Disease. Movement Disorders 37, 479–489. URL: https://onlinelibrary.wiley.com/doi/10.1002/mds.28945, doi:10.1002/mds.28945.

Marek, K., Chowdhury, S., Siderowf, A., Lasch, S., Coffey, C.S., Caspell-Garcia, C., Simuni, T., Jennings, D., Tanner, C.M., Trojanowski, J.Q., Shaw, L.M., Seibyl, J., Schuff, N., Singleton, A., Kieburtz, K., Toga, A.W., Mollenhauer, B., Galasko, D., Chahine, L.M., Weintraub, D., Foroud, T., Tosun-Turgut, D., Poston, K., Arnedo, V., Frasier, M., Sherer, T., Bressman, S., Merchant, M., Poewe, W., Kopil, C., Naito, A., Dorsey, R., Casaceli, C., Daegele,

N., Albani, J., Uribe, L., Foster, E., Long, J., Seedorff, N., Crawford, K., Smith, D., Casalin, P., Malferrari, G., Halter, C., Heathers, L., Russell, D., Factor, S., Hogarth, P., Amara, A., Hauser, R., Jankovic, J., Stern, M., Hu, S.C., Todd, G., Saunders-Pullman, R., Richard, I., Saint-Hilaire, H., Seppi, K., Shill, H., Fernandez, H., Trenkwalder, C., Oertel, W., Berg, D., Brockman, K., Wurster, I., Rosenthal, L., Tai, Y., Pavese, N., Barone, P., Isaacson, S., Espay, A., Rowe, D., Brandabur, M., Tetrud, J., Liang, G., Iranzo, A., Tolosa, E., Marder, K., Sanchez, M., Stefanis, L., Marti, M., Martinez, J., Corvol, J.C., Assly, O., Brillman, S., Giladi, N., Smejdir, D., Pelaggi, J., Kausar, F., Rees, L., Sommerfield, B., Cresswell, M., Blair, C., Williams, K., Zimmerman, G., Guthrie, S., Rawlins, A., Donharl, L., Hunter, C., Tran, B., Darin, A., Venkov, H., Thomas, C.A., James, R., Heim, B., Deritis, P., Sprenger, F., Raymond, D., Willeke, D., Obradov, Z., Mule, J., Monahan, N., Gauss, K., Fontaine, D., Szpak, D., McCoy, A., Dunlop, B., Payne, L., Ainscough, S., Carvajal, L., Silverstein, R., Espay, K., Ranola, M., Rezola, E., Santana, H., Stamelou, M., Garrido, A., Carvalho, S., Kristiansen, G., Specketer, K., Mirlman, A., Facheris, M., Soares, H., Mintun, A., Cedarbaum, J., Taylor, P., Jennings, D., Slieker, L., McBride, B., Watson, C., Montagut, E., Sheikh, Z., Bingol, B., Forrat, R., Sardi, P., Fischer, T., Reith, D., Egebjerg, J., Larsen, L., Breysse, N., Meulien, D., Saba, B., Kiyasova, V., Min, C., McAvoy, T., Umek, R., Iredale, P., Edgerton, J., Santi, D., Czech, C., Boess, F., Sevigny, J., Kremer, T., Grachev, I., Merchant, K., Avbersek, A., Muglia, P., Stewart, A., Prashad, R., and, J.T., 2018. The parkinson's progression markers initiative (PPMI) – establishing a PD biomarker cohort. Annals of Clinical and Translational Neurology 5, 1460–1477. URL: https://doi.org/10.1002/acn3.644, doi:10.1002/acn3.644.

Marquand, A.F., Kia, S.M., Zabihi, M., Wolfers, T., Buitelaar, J.K., Beckmann, C.F., 2019. Conceptualizing mental disorders as deviations from normative functioning. Molecular Psychiatry 24, 1415–1424. URL: https://doi.org/10.1038/s41380-019-0441-1, doi:10.1038/s41380-019-0441-1.

Meder, D., Herz, D.M., Rowe, J.B., Lehéricy, S., Siebner, H.R., 2019. The role of dopamine in the brain - lessons learned from Parkinson's disease. NeuroImage 190, 79–93. URL: https://linkinghub.elsevier.com/retrieve/pii/S1053811918320925, doi:10.1016/j.neuroimage.2018.11.021.

Muñoz-Ramírez, V., Kmetzsch, V., Forbes, F., Meoni, S., Moro, E., Dojat, M., 2022. Subtle anomaly detection: Application to brain MRI analysis of de novo Parkinsonian patients. Artificial Intelligence in Medicine 125, 102251. URL: https://linkinghub.elsevier.com/retrieve/pii/S0933365722000161, doi:10.1016/j.artmed.2022.102251.

Noroozi, M., Favaro, P., 2017. Unsupervised learning of visual representations by solving jigsaw puzzles. doi:https://doi.org/10.48550/arXiv.1603.09246.

van den Oord, A., Li, Y., Vinyals, O., 2019. Representation learning with contrastive predictive coding. doi:https://doi.org/10.48550/arXiv.1807.03748.

Oord, A.v.d., Vinyals, O., Kavukcuoglu, K., 2017. Neural discrete representation learning. URL: https://arxiv.org/abs/1711.00937, doi:10.48550/ARXIV.1711.00937.

Pinaya, W.H.L., Scarpazza, C., Garcia-Dias, R., Vieira, S., Baecker, L., F da Costa, P., Redolfi, A., Frisoni, G.B., Pievani, M., Calhoun, V.D., Sato, J.R., Mechelli, A., 2021a. Using normative modelling to detect disease progression in mild cognitive impairment and Alzheimer's disease in a cross-sectional multicohort study. Scientific Reports 11, 15746. URL: https://www.nature.com/articles/s41598-021-95098-0, doi:10.1038/s41598-021-95098-0.

Pinaya, W.H.L., Tudosiu, P.D., Gray, R., Rees, G., Nachev, P., Ourselin, S., Cardoso, M.J., 2021b. Unsupervised Brain Anomaly Detection and Segmentation with Transformers. URL: http://arxiv.org/abs/2102.11650. arXiv:2102.11650 [cs, eess, q-bio].

Poston, K.L., Ua Cruadhlaoich, M.A.I., Santoso, L.F., Bernstein, J.D., Liu, T., Wang, Y., Rutt, B., Kerchner, G.A., Zeineh, M.M., 2020. Substantia Nigra Volume Dissociates Bradykinesia and Rigidity from Tremor in Parkinson's Disease: A 7 Tesla Imaging Study. Journal of Parkinson's Disease 10, 591–604. doi:10.3233/JPD-191890.

Renieblas, G.P., Nogués, A.T., González, A.M., Gómez-Leon, N., del Castillo, E.G., 2017. Structural similarity index family for image quality assessment in radiological images. Journal of Medical Imaging 4, 035501. URL: https://doi.org/10.1117/1.jmi.4.3.035501, doi:10.1117/1.jmi.4.3.035501.

Rizek, P., Kumar, N., Jog, M.S., 2016. An update on the diagnosis and treatment of parkinson disease. Canadian Medical Association Journal 188, 1157–1165. URL: https://doi.org/10.1503/cmaj.151179, doi:10.1503/cmaj.151179.

Rutherford, S., Kia, S.M., Wolfers, T., Fraza, C., Zabihi, M., Dinga, R., Berthet, P., Worker, A., Verdi, S., Ruhe, H.G., Beckmann, C.F., Marquand, A.F., 2022. The normative modeling framework for computational psychiatry. Nature Protocols 17, 1711–1734. URL: https://www.nature.com/articles/s41596-022-00696-5, doi:10.1038/s41596-022-00696-5.

Shinde, S., Prasad, S., Saboo, Y., Kaushick, R., Saini, J., Pal, P.K., Ingalhalikar, M., 2019. Predictive markers for parkinson's disease using deep neural nets on neuromelanin sensitive MRI. NeuroImage: Clinical 22, 101748. URL: https://doi.org/10.1016/j.nicl.2019.101748, doi:10.1016/j.nicl.2019.101748.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M., 2017. Smoothgrad: removing noise by adding noise. doi:https://doi.org/10.48550/arxiv.1706.03825.

Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic attribution for deep networks. URL: https://arxiv.org/abs/1703.01365, doi:10.48550/ARXIV.1703.01365.

Tabelow, K., Balteau, E., Ashburner, J., Callaghan, M.F., Draganski, B., Helms, G., Kherif, F., Leutritz, T., Lutti, A., Phillips, C., Reimer, E., Ruthotto, L., Seif, M., Weiskopf, N., Ziegler, G., Mohammadi, S., 2019. hMRI – A toolbox for quantitative MRI in neuroscience and clinical research. NeuroImage 194, 191–210. URL: https://linkinghub.elsevier.com/retrieve/pii/S1053811919300291, doi:10.1016/j.neuroimage.2019.01.029.

Taleb, A., Loetzsch, W., Danz, N., Severin, J., Gaertner, T., Bergner, B., Lippert, C., 2020. 3d self-supervised methods for medical imaging, in: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 18158–18172. URL: https://proceedings.neurips.cc/paper/2020/file/d2dc6368837861b42020ee72b0896182-Paper.pdf.

Trujillo, P., Summers, P.E., Ferrari, E., Zucca, F.A., Sturini, M., Mainardi, L.T., Cerutti, S., Smith, A.K., Smith, S.A., Zecca, L., Costa, A., 2017. Contrast mechanisms associated with neuromelanin-MRI: Neuromelanin-MRI Contrast. Magnetic Resonance in Medicine 78, 1790–1800. URL: https://onlinelibrary.wiley.com/doi/10.1002/mrm.26584, doi:10.1002/mrm.26584.

Tschuchnig, M.E., Gadermayr, M., 2021. Anomaly detection in medical imaging – a mini review URL: https://arxiv.org/abs/2108.11986, doi:10.48550/ARXIV.2108.11986.

Tudosiu, P.D., Pinaya, W.H.L., Graham, M.S., Borges, P., Fernandez, V., Yang, D., Appleyard, J., Novati, G., Mehra, D., Vella, M., Nachev, P., Ourselin, S., Cardoso, J., 2022. Morphology-preserving autoregressive 3d generative modelling of the brain. URL: https://arxiv.org/abs/2209.03177, doi:10.48550/ARXIV.2209.03177.

Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E., 2004. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing 13, 600–612. URL: https://doi.org/10.1109/tip.2003.819861, doi:10.1109/tip.2003.819861.

Weiskopf, N., Suckling, J., Williams, G., Correia, M.M., Inkster, B., Tait, R., Ooi, C., Bullmore, E.T., Lutti, A., 2013. Quantitative multi-parameter mapping of R1, PD*, MT, and R2* at 3T: a multi-center validation. Frontiers in Neuroscience 7. URL: http://journal.frontiersin.

org/article/10.3389/fnins.2013.00095/abstract, doi:10.3389/fnins.2013.00095.

Wenger, E., Polk, S.E., Kleemeyer, M.M., Weiskopf, N., Bodammer, N.C., Lindenberger, U., Brandmaier, A.M., 2021. Reliability of quantitative multiparameter maps is high for MT and PD but attenuated for R1 and R2* in healthy young adults. preprint. Neuroscience. URL: http://biorxiv.org/lookup/doi/10.1101/2021.11.10.467254, doi:10.1101/2021.11.10.467254.

Xiao, Y., Fonov, V., Bériault, S., Subaie, F.A., Chakravarty, M.M., Sadikot, A.F., Pike, G.B., Collins, D.L., 2014. Multi-contrast unbiased MRI atlas of a parkinson's disease population. International Journal of Computer Assisted Radiology and Surgery 10, 329–341. URL: https://doi.org/10.1007/s11548-014-1068-y, doi:10.1007/s11548-014-1068-y.

Zeiler, M.D., Fergus, R., 2013. Visualizing and understanding convolutional networks. URL: https://arxiv.org/abs/1311.2901, doi:10.48550/ARXIV.1311.2901.

Zucca, F.A., Segura-Aguilar, J., Ferrari, E., Muñoz, P., Paris, I., Sulzer, D., Sarna, T., Casella, L., Zecca, L., 2017. Interactions of iron, dopamine and neuromelanin pathways in brain aging and Parkinson's disease. Progress in Neurobiology 155, 96–119. URL: https://linkinghub.elsevier.com/retrieve/pii/S030100821500101X, doi:10.1016/j.pneurobio.2015.09.012.

Appendix A. Extra figures

PD25 atlas		Synthseg labels			
Label	Nuclei	Label	ROI	Label	ROI
1	Left red nucleus	2	Left cerebral white matter	26	Left accumbens area
3	Left substantia nigra	3	Left cerebral cortex	28	Left ventral DC
5	Left subthalamic nucleus	4	Left lateral ventricle	41	Right cerebral white matter
7	Left caudate	5	Left inferior lateral ventricle	42	Right cerebral cortex
9	Left putamen	7	Left cerebellum white matter	43	Right lateral ventricle
11	Left globus pallidus externa	8	Left cerebellum cortex	44	Right inferior lateral ventricle
13	Left globus pallidus interna	10	Left thalamus	46	Right cerebellum white matter
15	Left thalamus	11	Left caudate	47	Right cerebellum cortex
2	Right red nucleus	12	Left putamen	49	Right thalamus
4	Right substantia nigra	13	Left pallidum	50	Right caudate
6	Right subthalamic nucleus	14	3rd ventricle	51	Right putamen
8	Right caudate	15	4th ventricle	52	Right pallidum
10	Right putamen	16	Brain-stem	53	Right hippocampus
12	Right globus pallidus externa	17	Left hippocampus	54	Right amygdala
14	Right globus pallidus interna	18	Left amygdala	58	Right accumbens area
16	Right thalamus	24	CSF	60	Right ventral DC

Table A.4: Atlas labels

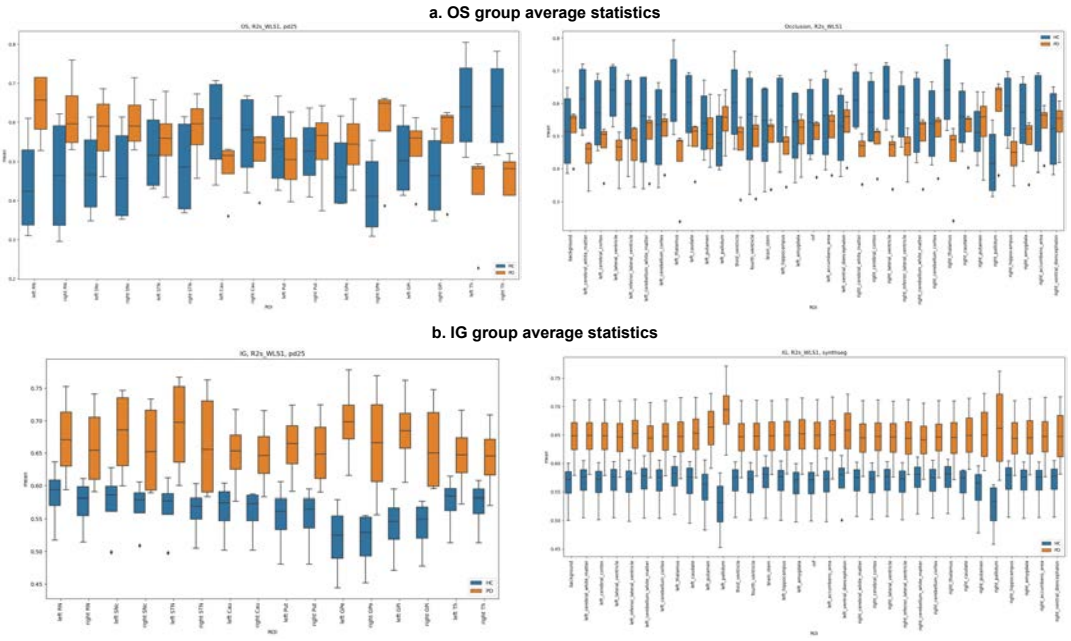


Figure A.12: Group average statistics for R2* experiment, using normalized OS and IG maps

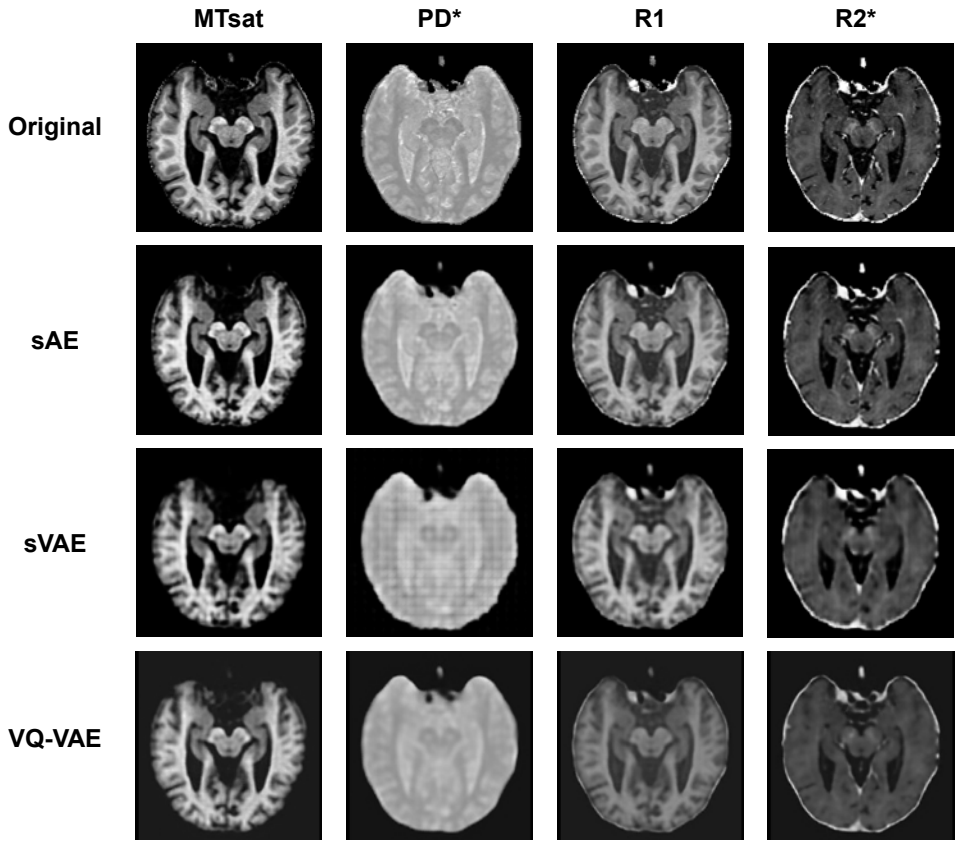
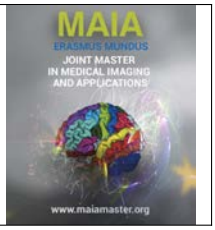


Figure A.13: Reconstruction examples for all qMRI maps and the three different model architectures.

Medical Imaging and Applications

Master Thesis, June 2023



Revisiting Long-Tailed Learning From a Free-Lunch Perspective

Marawan Elbatel[a,b], Robert Martí[a], Xiaomeng Li[b]

[a] *Institute of Computer Vision and Robotics, University of Girona, Girona, Spain*
[b] *The Hong Kong University of Science and Technology*

Abstract

Long-tailed learning has the potential to provide significant benefits in various real-world applications, especially within the medical field where certain diseases and conditions, such as skin lesions classification and gastrointestinal recognition, exhibit a long-tailed distribution. Existing methods primarily rely on domain-specific optimization objectives, hindering their ability to effectively handle rare diseases due to lacking a generalizable feature representation. In this thesis, we revisit long-tailed learning by utilizing publicly available pre-trained models, often called "free lunch models". Specifically, we propose effective knowledge distillation (EKD) to distill publicly available pre-trained models to smaller target medical models in centralized and decentralized settings. For a centralized setting, we present *Fourier Prompted Knowledge Distillation (FoPro-KD)* unleashing the power of frequency patterns learned from frozen publicly available pre-trained models to enhance their transferability and compression. For decentralized training, specifically federated learning, we investigate the learning dynamics of our proposed EKD in local clients and present *FedFree*, a framework enabling federated long-tailed learning, providing valuable insights and unveiling significant findings that can be derived from the utilization of pre-trained models. We evaluate the effectiveness of our proposed frameworks on two long-tailed learning benchmarks, gastrointestinal and skin lesion recognition tasks. The experimental results demonstrate the favorable performance achieved by both *FoPro-KD* and *FedFree* in their respective settings for long-tailed medical imaging recognition.

Keywords: Long-Tailed Learning, Knowledge Distillation, Federated Learning

1. Introduction

Long-tailed distributions, characterized by severe class imbalance where majority classes significantly outnumber minority classes, are common in many medical imaging tasks, such as skin-lesion classification and gastrointestinal image recognition (Borgli et al., 2019; Combalia et al., 2019; Tschandl et al., 2018). While convolutional neural networks (CNNs) have demonstrated remarkable performance in medical image classification, their application can be limited in the presence of scarce labeled medical image datasets, particularly in long-tailed datasets with rare diseases. To address this challenge, transfer learning has emerged as a promising approach, aiming to fine-tune pre-trained models trained on natural images for improved performance on medical image datasets. However, an important consideration in transfer learning is to develop an efficient technique that not only preserves the generalization capabilities of large pre-trained models but also ensures compactness for practical deployment.

Publicly available pre-trained models, such as MoCo (He et al., 2020), BYOL (Grill et al., 2020), CLIP (Radford et al., 2021), and DINO (Oquab et al., 2023), have attracted considerable attention in the medical imaging community due to their promising generalization capabilities as "free lunch" models (Ding et al., 2022). However, these pre-trained models' extensive complexity and significant computational resource requirements can limit their applicability in clinical settings in low infrastructure, point-of-care testing, and edge devices. Moreover, fine-tuning (FT) these models on smaller, long-tailed medical image datasets can distort the generalizability of these models (Kumar et al., 2022). Therefore, developing an effective transfer learning approach is highly demanded to leverage the generalization capabilities of large pre-trained models while maintaining performance on the target task.

In this thesis, we propose effective knowledge distillation (EKD) to enhance the transfer of publicly available pre-trained models, known as "free lunch models," to smaller target medical imaging models. First, we investigate the inherent characteristics of these pre-trained models in a centralized setting, particularly their preferred input frequencies and semantics. Then, we extend our investigation to decentralized training scenarios.

Recently, Yu et al. (2023) quantified the frequency bias in neural networks and proposed a method for guiding the network to tune its frequency by utilizing a Sobolev norm that expands the L2 norm. Although their approach was limited to Neural Tangent Kernels (NTK) and focused on quantifying the frequency bias on a broad frequency basis, their work inspired us to explore and exploit these patterns from publicly available pre-trained models conditioned on a target medical dataset to improve the representation learning for rare disease classification. Pre-trained models encode frequency patterns through their convolutional and pooling operations during pre-training. Each filter in the convolutional layer acts as a frequency filter, capturing distinct patterns in the input data while pooling operations, further amplify or attenuate these patterns. This frequency-dependent behavior can introduce biases in the model, making it more sensitive to certain frequency patterns and less sensitive to others, which may not align with the frequency characteristics of target medical data. To this end, we propose *FoPro-KD (Fourier-prompted Knowledge Distillation)* for centralized training. FoPro-KD explores and exploits the learned frequency patterns from publicly available pre-trained models conditioned on a target medical dataset to improve the representation transfer for rare disease classification as depicted in Figure 1.

The applicability of large pre-trained models in centralized scenarios is promising; however, the privacy-focused and decentralized nature of medical imaging data, coupled with its inherent data heterogeneity, presents a challenge in training a robust global model. To address the challenges posed by data decentralization, we extend our proposed effective knowledge distillation (EKD) to a decentralized training scenario. Specifically, we notice that leveraging publicly available pre-trained models locally at each client can work as consistent reference frames for measuring local bias. Based on this insight, we introduce *FedFree (Federated learning via leveraging free lunch models)*, a framework incorporating a novel dynamic long-tailed model aggregation (DLMA). DLMA captures inter-client intra-class variations and facilities robust global model aggregation.

The main contributions of this work can be summarized as the following:

- We demonstrate that effective knowledge distillation (EKD) from frozen pre-trained models on natural images to a target smaller medical imaging model can be just as effective as traditional long-tailed methods, thanks to their generalization capabilities.
- For centralized training, we introduce a novel framework called FoPro-KD to improve the transferability of pre-trained models to smaller medical imaging models. Specifically, we generate targeted perturbations as fourier spectral prompts that further improve the distillation process.
- We explore the learning dynamics of our proposed EKD in a decentralized setting, leading to the development of FedFree. Fedfree encompasses a novel dynamic long-tailed aggregation method to address the challenges posed by inter-client intra-class variations, which can impede effective representation learning in decentralized training.
- We evaluate our frameworks on two challenging long-tailed datasets, the skin lesion classification and a more challenging gastrointestinal image recognition testbed. FoPro-KD and FedFree surpass the state-of-the-art methods in both datasets in centralized and decentralized settings, respectively.

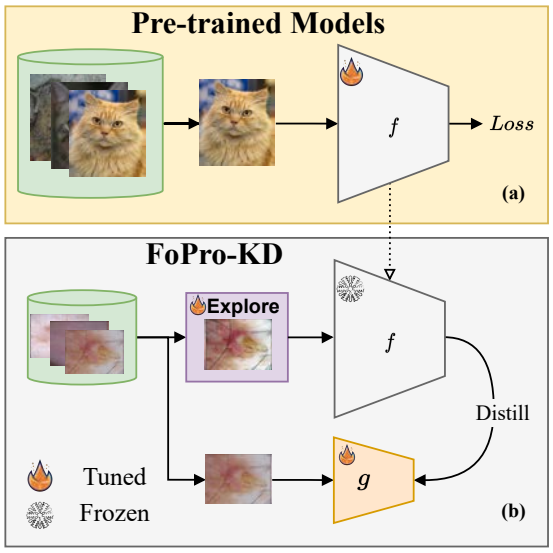


Figure 1: (a) The Free Lunch model assumes specific frequency patterns in input data. (b) Our FoPro-KD approach explicitly queries the model to identify meaningful frequency patterns for distillation.

2. Related Work

In this section, we review the literature related to transfer learning with prompt tuning, adversarial domain adaptation, long-tail, and federated long-tailed learning.

2.1. Transfer Learning

In recent years, transfer learning and fine-tuning have been extensively studied in the literature, with a focus on adapting the feature extractor to fit the target task. However, such approaches can deviate from pre-trained features, resulting in a trade-off between the performance of the majority class (in-distribution or IID) and the rare class (out-of-distribution or OOD). To mitigate similar tradeoffs on IID and OOD datasets, Kumar et al. (2022) proposed a simple variant of initializing the head with a linear probed version followed by full fine-tuning. Nevertheless, these methods can suffer from deviating semantics and extreme overfitting on long-tailed problems when fully fine-tuning large pre-trained models. Prompt tuning arises in vision to address these issues for efficiently fine-tuning large models in vision tasks, similar to natural language processing (NLP). Jia et al. (2022) proposed Vision Prompt Tuning (VPT), which adds prompts to vision transformers and exploits the transformer's location-invariant features for effective fine-tuning. Similar to NLP prompt tuning, Dong et al. (2023) explored the use of prompt learning for the effective transfer of pre-trained vision transformers for long-tail natural image classification. These methods are specially tailored to vision transforms similar to NLP, failing to find an efficient prompt for transforming the knowledge of CNN vision-pre-trained models, which are important for medical imaging classification. Recent studies have shown that DNNs rely on high-frequency patterns, which are typically ignored by radiologists for output representations (Makino et al., 2020). Morever, Bai et al. (2022) found that a CNN teacher can benefit vision transformers to fit high-frequency components and proposed HAT to adversarially augment images' high-frequency components towards improving vision transformers generalization capabilities. Prompt tuning for CNN models can be related to the literature on adversarial learning and domain adaptation.

2.2. Adversarial learning

Adversarial learning has emerged as a popular approach for domain adaptation (DA) and domain generalization (DG). To achieve DA, Huang et al. (2021) proposed a method that generates adversarial examples from the source dataset and fine-tunes the model on the target dataset using both adversarial and clean examples. Similarly, Kim et al. (2023) modeled DG as DA to adversarially generate worst-case targets from the source dataset. Chen et al. (2022a) proposed MaxStyle as an adversarial realistic data augmentation utilizing an auxiliary image decoder for robust medical image segmentation. For source-free unsupervised domain adaptation (SFUDA), Hu et al. (2022) proposed to learn a domain-aware prompt adversarially for a UNet-based model. More recently, Wang et al. (2023), inspired by Fourier style mining (Yang et al., 2022), proposed to learn a low-frequency Fourier visual prompt for SFUDA that excelled in segmentation performance. However, all these methods are restricted to source and target datasets trained for the same closed-set task and often rely on increasing noise to synthesize adversarial examples in DG or on bridging the gap between datasets in DA. In addition, their approaches do not explicitly leverage the frequency patterns captured by pre-trained models on natural images, which can aid in representational learning, especially for long-tailed datasets.

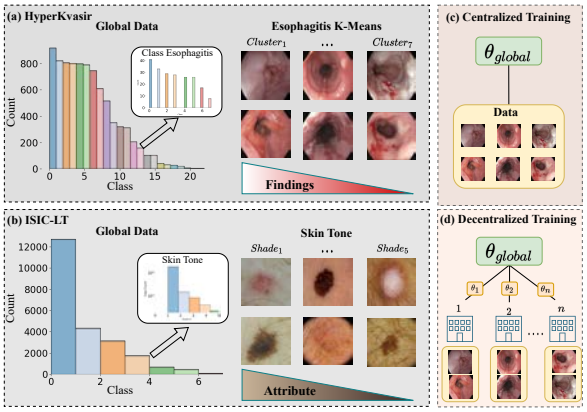


Figure 2: (a) The HyperKvasir Dataset. (b) The ISIC-LT dataset. (c) Centralized training, (d) Decentralized Training.

2.3. Long-Tail Learning

The severe class imbalance in long-tailed (LT) learning poses challenges for training accurate models, and various approaches have been proposed to address this issue, including data augmentation techniques, re-sampling and re-weighting schemes, and curriculum-based methods. Data augmentation techniques aim to regularize the model by incorporating regularization techniques to enhance the model's representations. For instance, Zhang et al. (2018) proposed MixUp, which utilizes linear interpolation between two images with soft labeling to provide information augmentation for regularization during training. Chen et al. (2021b) introduced Amplitude-Phase Recombination (APR), which focuses on swapping amplitudes between images to reduce sensitivity to amplitude shifts.

These data augmentations need to be coupled with a balancing scheme to account for the extreme class imbalance in LT datasets. Galdran et al. (2021) proposed Balanced-Mixup, a simple variant of MixUp using class conditional sampling that has compelling capabilities for highly imbalanced medical image classification. Nevertheless, data augmentation methods do not account for the label distribution shift that can arise over the test set. Class balancing loss (CB) (Cui et al., 2019), Label distribution margin (LDAM) (Cao et al., 2019), and balanced-softmax (BSM) (Ren et al., 2020) was proposed as modified re-weighting strategies for training models for long-tailed learning. However, these

methods often have limitations, such as not effectively addressing the extreme bias from head classes. To address such bias, Kang et al. (2020) found that the classifier is the major bottleneck for the head classes bias in long-tail learning and proposed a two-stage learning approach that decouples the feature extractor representations from the classifier through a plug-in classifier re-training (cRT). Despite the performance improvements achieved by cRT in various long-tailed methods, it fails to address the issue of intra-class imbalance that can limit effective representation extraction (Zhao et al., 2021).

To demonstrate this, we present the imbalance attributes within the HyperKvasir (Borgli et al., 2019) dataset in Figure 2 (a). The dataset exhibits instances with different findings for the same class, such as trachealization, varices, erosion with leukoplakia. We cluster the features of the free-lunch model and visualize these attributes specifically for the Esophagitis class. Additionally, the ISIC-LT dataset presents extreme class imbalance across different skin tones (Bevan and Atapour-Abarghouei, 2022), as shown in Figure 2 (b).

To tackle the intra-class imbalance, a previous study by Tang et al. (2022) proposed invariant feature learning (IFL) through dual environment learning and resampling techniques. On the other hand, methods based on curriculum learning, requiring a pre-training stage on the target dataset to extract meaningful representation followed by utilizing these representations, have achieved state-of-the-art (SOTA) performance for long-tailed learning. For example, Zhang et al. (2023) achieved SOTA in multiple long-tailed datasets by a two-stage framework. First, by pre-training a teacher model on the target dataset to capture the target dataset representations, followed by a balanced knowledge distillation (BKD) to guide a student model. However, all the aforementioned methods have not utilized the generalization capabilities of publicly available pre-trained models known for their generalizable representations, as they focus more on the problem on a narrow knowledge extraction basis from the target dataset, whereas pre-training and the knowledge gained from natural images have achieved compelling performance in medical imaging as "free lunch models" (Ding et al., 2022).

In our work, we re-visit long-tailed learning in medical imaging from a free lunch perspective. We demonstrate that the generalizable features from publicly available pre-trained models on natural images can be comparable to different long-tail methods without additional pre-training or fine-tuning of these free lunch models. In addition, we find that these free lunch models have a preferred frequency basis (i.e. styles) for their input that can restrict their distillation in many tasks. To address such preferred styles, we propose to explore these preferred styles through effective prompting on a frequency basis. By exploring the pre-trained models' frequency patterns and iteratively distilling such knowledge, we can recycle and compress these pre-trained models with no additional training to the target medical task, our approach can be easily utilized with different long-tailed learning schemes as a free lunch distillation, achieving SOTA on multiple long-tailed medical imaging datasets. Addressing long-tailed challenges in the centralized setting as in Figure 2 (c), our method proves to be effective. However, the task becomes even more daunting when the dataset is decentralized across different clients, as depicted in Figure 2 (d).

2.4. Federated Learning

Federated learning (FL) has emerged as a way to train models with this decentralized data while preserving privacy. However, this decentralization has led to a degradation in the performance of both generic and personalized models due to issues with data heterogeneity. This issue is especially critical when dealing with long-tailed datasets. With FedAvg (McMahan et al., 2017) as the main baseline, multiple works propose to improve the model's generic performance under data heterogeneity (Li et al., 2021, 2020; Mendieta et al., 2022). While these methods have been successful in achieving positive results while assuming a balanced global data distribution, they have struggled when dealing with extreme data heterogeneity, particularly in the case of long-tailed (LT) datasets. Although there have been some methods proposed to address the imbalanced setting (Liu et al., 2021; Mu et al., 2021), these methods shared local features (such as correlation matrix) among clients, which may raise privacy concerns for the clients. Additionally, the issue of label distribution skewness has been addressed in the context of federated learning (Oh et al., 2022; Zhang et al., 2022). While these methods have shown promising results by adjusting the local class distributions, they do not explicitly address the inherent extreme label distribution skewness present in long-tailed learning. This specific characteristic of long-tailed learning poses unique challenges that need to be explicitly accounted for in the context of federated learning.

The issue of federated long-tailed (Fed-LT) was initially addressed by (Shang et al., 2022). The authors proposed CReFF to handle synthetically generated LT natural image datasets. Their approach involved retraining a new classifier by leveraging learnable features on the server at the cost of uploading clients' gradients over their local distribution to the server. In a recent study, Chen et al. (2022c) showed that despite its substantial server overhead for communication and computation, CReff only provides a minor improvement compared to traditional FL methods. Different approaches have also been proposed to regularize the local client training procedure to address the Fed-LT challenge. For instance, Shuai et al. (2022) incorporated knowledge distillation (KD) from the global model, inspired by (Li

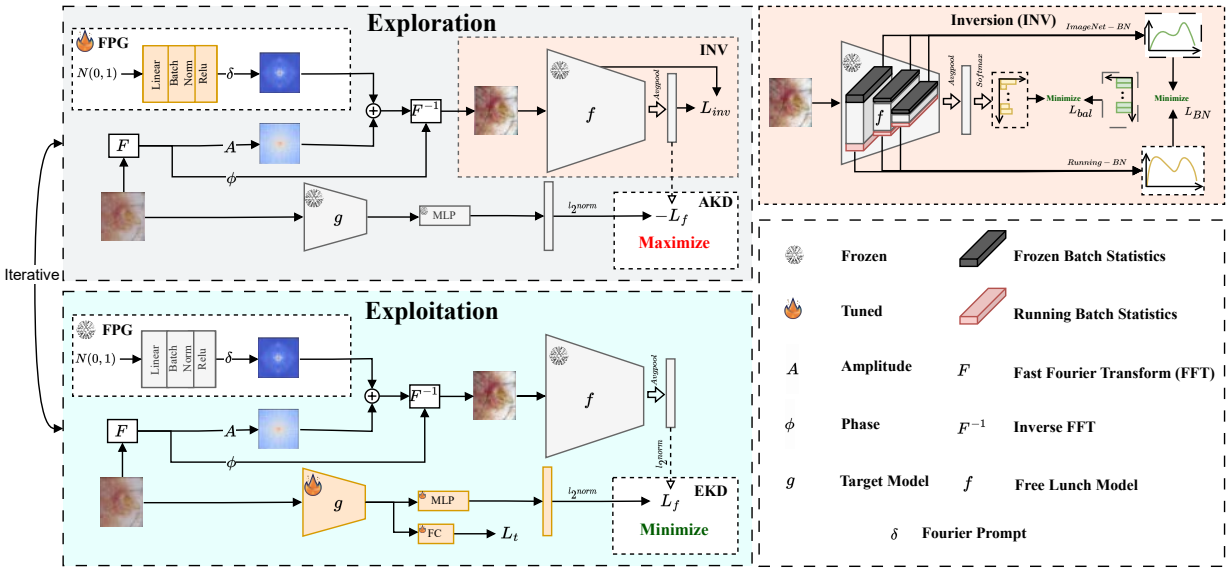


Figure 3: Our proposed FoPro-KD framework has two phases: exploration and exploitation. In the exploration phase, the FPG generates Fourier prompts to capture frequency patterns of the frozen pre-trained model f. In the exploitation phase, the proposed effective knowledge distillation (EKD) module distills the knowledge from f into the target model g, guided by the Fourier prompt generator (FPG). Our framework can iteratively alternate between the exploration and exploitation phases using adversarial knowledge distillation (AKD) to enhance representation distillation and learning efficiency of g.

et al., 2021), for missing classes and applied local regularization (Pereyra et al., 2017) for the majority classes. Their method is specifically designed for local training. However, when coupled with standard FedAvg, different long-tailed methods like Balanced-Softmax (Ren et al., 2020) demonstrate substantial performance improvements (Wicaksana et al., 2023). More recently, several approaches have emerged to address the challenges of Federated long-tailed learning (Li et al., 2023; Wicaksana et al., 2023; Wu et al., 2022; Yang et al., 2023). These methods tackle the issues of label skewness and class bias by adopting decoupled training, separating the classifier and the feature extractor. The rationale behind these methods is rooted in the understanding that the classifier plays a significant role in label skewness and class bias. By decoupling it from the feature extractor, these challenges can be effectively mitigated (Zhao et al., 2021).

A notable limitation of these approaches is their inability to consider the inter-client intra-class variations that emerge due to the federated long-tailed distribution. Moreover, these methods heavily rely on aggregated representations from the global model, which restricts further performance improvements in the absence of local generalizable representations.

To this end, we study the learning behavior of our proposed EKD in a decentralized setting to propose **FedFree** towards robust federated learning via leveraging free lunch models. Unlike previous methods that focus on local training (Shuai et al., 2022; Zhang et al., 2022), we study the inter-client intra-class variations with our proposed effective knowledge distillation (EKD) to identify clients that are not well captured by the global model due to Fed-LT behavior. Based on this, we derive a dynamic long-tailed-aware model aggregation (DLMA) that gives higher weights to client-specific models, thereby capturing their local variance and contributing to a more generalized global model.

3. Method

This section describes our effective utilization of "free lunch models" in centralized and decentralized training settings. We first introduce our framework for centralized training, known as FoPro-KD. We then present our proposed framework for decentralized training, known as FedFree.

*3.1. **FoPro-KD** (Centralized Framework)*

Figure 3 shows the framework for our proposed FoPro-KD. The training of FoPro-KD consists of two stages: an exploration stage and an exploitation stage. In the exploration stage, we train one linear layer as a Fourier Prompt Generator (FPG) to generate Fourier amplitude spectral prompts, δ, conditional on our target medical data, allowing us to explore the representations of the free lunch model, f, by explicitly asking what frequency patterns on the input lead to meaningful representations. This is done while freezing f, pre-trained on natural imaging dataset (ex: ImageNet (He et al., 2020)). In the exploitation stage, we effectively distill these generalizable representations to a smaller target medical imaging model, g through our proposed Effective Knowledge Distillation (EKD). To make the Fourier

prompts more diverse while being representative of f, we perform multiple iterations of the exploration and exploitation stages by an Adversarial Knowledge Distillation (AKD). This allows us to effectively exploit the generalization capabilities of large pre-trained models and compress them into smaller student networks that are useful for practical medical imaging deployment in a clinical setting.

3.1.1. Fourier Prompt Generation

To attain optimal representational transfer, publicly available pre-trained models necessitate input data that closely align with their preferences. In this regard, training a conditional generative adversarial network (CGAN) (Mirza and Osindero, 2014) to guide the target dataset towards these preferences can substantially modify the semantics of the dataset. As shown in Figure 4, training a CGAN with deep inversion causes modification in the semantics of the target dataset in the highly informative regions conditional on the semantics of the pre-training dataset, ImageNet (Deng et al., 2009).

Recent research by Yu et al. (2023) has shown theoretically that neural networks can be sensitive to certain frequencies without explicitly considering the frequency patterns captured during pre-training deep neural networks (DNNs). Therefore, we aim to explore this frequency-dependent behavior of CNNs and enable frozen pre-trained models to output representations through prompting on a frequency basis, which is facilitated by our proposed Fourier Prompt Generator (FPG).

FPG employs a random noise vector, z, to generate a three-dimensional Fourier amplitude prompts, $\delta = FPG(z)$, one for each channel respectively, enabling the modification of the target dataset by emphasizing or suppressing specific frequency patterns preferred and captured by "free lunch models" on the source natural images dataset. Although these preferred patterns relied on the deep learning dynamics of the "free lunch models", the FPG can be trained to unleash such patterns and generate Fourier prompts that are the preference of the "free lunch model" conditioned on our target medical dataset. This feature plays a critical role in effective knowledge distillation.

Let the Fourier decomposition of an image x be F(x), which consists of the amplitude A and phase ϕ components:

$$F(x) = A \odot e^{i\phi} \tag{1}$$

To interpolate the Fourier amplitude between the input image and the generated Fourier prompt, we use a mixing coefficient, denoted by α and sampled uniformly from 0 to 1, resulting in a new Fourier amplitude spectrum $\hat{A}$:

$$\hat{A}_{ij} = \alpha A_{ij} + (1 - \alpha)\delta_{ij} \tag{2}$$

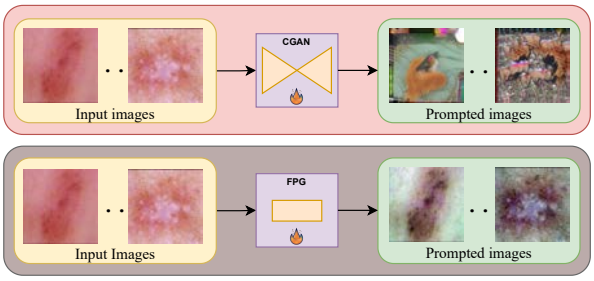


Figure 4: Using a conditional GAN (CGAN) to manipulate the input dataset changes the image semantics in highly informative regions compared to surpassing or amplifying certain frequencies in these regions with FPG.

where A_{ij} represents the Fourier amplitude of the input image, δ_{ij} represents the the generated Fourier prompt, and ij are the indices of the Fourier coefficients.

The modified Fourier coefficients are then transformed back using the inverse Fourier transform to generate the modified image, denoted by $\hat{x}$,

$$\hat{x} = F^{-1}(\hat{A} \odot e^{i\phi}) \tag{3}$$

where F^{-1} denotes the inverse Fourier transform.

We train the Fourier prompt generator in the exploration phase while freezing all other modules. Specifically, we feed $\hat{x}$ to the frozen pre-trained feature extractor, f, and utilize the batch regularization technique that was first introduced by Ye et al. (2020). This technique minimizes the divergence between the feature statistics, which include the mean and variance of the features, and the corresponding batch normalization statistics by assuming a Gaussian distribution:

$$\mathcal{L}_{\text{BN}}(x) = \sum_{l \in f} D\Big(N\big(\mu_l(\hat{x}), \sigma_l^2(\hat{x})\big) \Big| N\big(\mu_l, \sigma_l^2\big)\Big), \tag{4}$$

where D is the L2 divergence loss, $N(\mu_l(\hat{x}), \sigma_l^2(\hat{x}))$ is the feature statistics of the modified input batch $\hat{x}$, $N(\mu_l, \sigma_l^2)$ is the batch normalization statistics of the frozen model, f, and l indexes the layers of f.

To better capture the frozen pre-trained model's learned frequency patterns and avoid skewing in the learning of the Fourier Prompt Generator (FPG), we propose a regularization approach that encourages the synthesis of Fourier prompts with a more balanced distribution of activations across the final pre-classification features. This is achieved by maximizing the entropy of the free lunch model output towards a uniform distribution where each feature has an equal probability of being activated as

$$\mathcal{L}_{\text{bal}} = \sum_{i=1}^{C} p_i \log p_i \tag{5}$$

where C is the dimension of the final pre-classification features, and p_i is the i-th element of the softmax output

p of the frozen pre-trained model on the target modified data $\hat{x}$. This approach avoids bias towards any particular feature and promotes the generalization ability of the learned Fourier prompts.

The final inversion loss $\mathcal{L}_{inv}$ to train the FPG module is defined as the combination of the batch normalization loss, $\mathcal{L}_{BN}$, and the balancing loss, $\mathcal{L}_{bal}$, as

$$\mathcal{L}_{inv} = \mathcal{L}_{BN} + \mu\, \mathcal{L}_{bal} \tag{6}$$

where μ is the weighting factor for the balancing regularization.

Combining this balanced regularization term with the batch statistics losses, the generated Fourier prompts can exhibit higher entropy while being specific to the frozen pre-trained model's desired frequencies to better benefit the knowledge distillation. We apply a Hermitian constraint to ensure that the generated Fourier prompts produce valid Fourier amplitudes.

The exploration phase ensures that the Fourier generator produces styles consistent with the preferred frequency patterns of the free lunch model while avoiding overfitting specific styles.

Our training approach for the FPG can be seen as a deep inversion method in the literature of data-free knowledge distillation (Fang et al., 2021). However, our method is unique in the learnable and target objectives, in addition, conditioned on a cross-task target dataset, which makes it more challenging.

3.1.2. Exploitation with Effective Knowledge Distillation

Large pre-trained models available to the public possess remarkable generalization capabilities that can assist in the classification of rare diseases. It has been observed that performing linear probing on these models yields high-accuracy results on out-of-distribution (OOD) datasets. However, complete fine-tuning of these models may lead to distortion of these highly generalizable representations (Kumar et al., 2022). To this end, we propose Effective Knowledge Distillation (EKD), which aims to compress the generalization capabilities of the free lunch models while maintaining generalizable performance on the target data using a smaller model.

To achieve this, we utilize a small target model with a feature extractor $g(\cdot)$ to be trained on the target medical dataset, along with a large frozen publicly available pre-trained encoder $f(\cdot)$ (free lunch model). To compare the latent features of the target model with those of the free lunch model, we add a 2-layer MLP on top of the smaller target feature extractor, $g(\cdot)$.

To generate the necessary encodings for distillation, we sample an image x uniformly from the target dataset $\mathcal{D}$ and use prompt mixing following Equation (3) to obtain $\hat{x}$, while freezing FPG. This allows us to navigate the representation of f based on the styles and frequencies it was trained on.

From here, we generate two encodings: a *projection* $y = MLP(g(x))$ and a *target representation* $t = f(\hat{x})$ from our target network and the large frozen pre-trained network, respectively. We then L2-normalize both encodings and distill the information from the large pre-trained model to the smaller target model using a mean squared error loss as our distillation loss between both encodings, as

$$\mathcal{L}_f = 2 - 2 \cdot \langle y, t \rangle \cdot \tag{7}$$

While previous approaches (Chen et al., 2022b, 2021a) aims to reduce the performance gap between the teacher and student models on the same task, our proposed distillation loss is designed to narrow the generalization capabilities of free lunch models to a different task, which our student model is being trained on. This approach can act as an implicit regularization technique, leveraging the discriminative generalization capabilities of large pre-trained model features for the tail classes. Specifically, our approach encourages the $g(\cdot)$ to generalize well to the tail classes of the target task, which may be rare and difficult to identify without additional guidance. We denote the exploitation loss, $\mathcal{L}_{exploit}$, to minimize at each training step as:

$$\mathcal{L}_{exploit} = \mathcal{L}_t + \lambda_f\, \mathcal{L}_f, \tag{8}$$

where $\mathcal{L}_t$ is a balanced risk minimization (Ren et al., 2020).

3.1.3. Exploration with Adversarial Distillation

To ensure that the learned Fourier amplitudes become more diverse while being representative of the natural image styles, thus alleviating any representational mode collapse issue in distillation, we propose to further enhance the Fourier prompt generation by navigating the latent space of the free lunch model with an iterative adversarial loss.

To achieve this, we propose maximizing the proposed effective knowledge distillation (EKD) loss between the free lunch model and the target model for iterative exploration. The final exploration loss, $\mathcal{L}_{explore}$, to be optimized is given by:

$$\mathcal{L}_{explore} = -\gamma \lambda f \mathcal{L}_f + \mathcal{L}_{inv} \tag{9}$$

Here we maximize the similarity between the free-lunch model and the target model, as described in Equation (7). This adversarial loss is weighted by a hyperparameter γ, which determines the strength of the adversarial training. Unlike standard adversarial training, we aim to explore the free-lunch model, so we set γ between 0 and 1, with an upper bound of the exploitation distillation factor λ_f. This is similar to the training of generative adversarial networks (GANs) (Goodfellow et al., 2014). To this end, we choose a value of $\gamma = 0.3$ and provide an ablation study to validate our

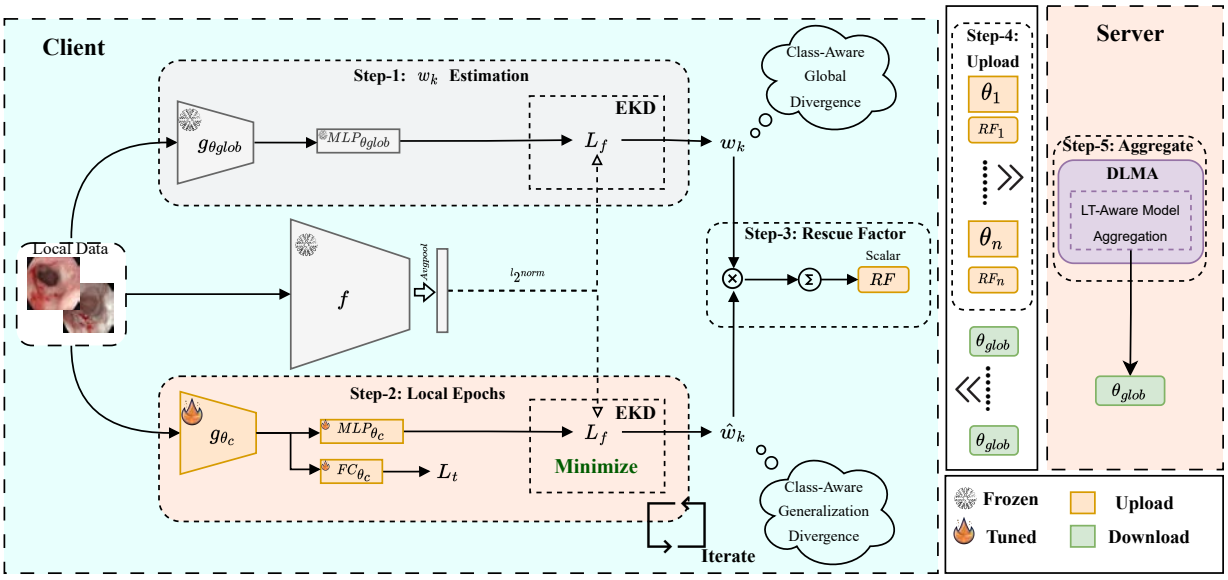


Figure 5: Overall scheme of FedFree framework.

choice. $\mathcal{L}_{inv}$ ensures that the prompts generated by the Fourier prompt generator accurately represent the pre-trained model.

3.2. ***FedFree*** *(Decentralized Training)*

While FoPro-KD leverages pre-trained models for centralized training, the significant data heterogeneity under data decentralization makes it more challenging to train a robust global model. In this regard, we investigate the application of "free lunch models" in federated learning.

Figure 5 shows the overview of our FedFree framework. Each local client is provided with a publicly self-supervised pre-trained model (e.g., MoCo-RN50 (He et al., 2020)) that is not involved in the training or communication process of the federated learning framework. These "free lunch models" do not increase communication costs while ensuring each client can access discriminative unbiased representation from the same consistent model. With n local clients and one global server, our FedFree performs the following steps in each round: (1) Each client receives the global model to update its local model and estimate global class-aware $\mathcal{L}_f$ using Equation (7); (2) Each client trains its local model to minimize Equation (8) while estimating its local class-aware $\mathcal{L}_f$; (3) Each client calculates a rescue scalar; (4) Client uploads the parameters of its local model and RF to the server; (5) Using our proposed DLMA, the server aggregates a new model from the parameters of the received client models, weighted by RF;

3.2.1. *EKD in FL setting*

"Free lunch models" uniquely offer a consistent discriminative distribution for all clients, facilitating the alignment of diverse client learning processes with a shared discriminative distribution at the local level. In our study, we explore the application of our proposed EKD (Equation (7)) in a decentralized setting. To ensure a fair comparison with other federated learning methods, we exclude the FPG from our decentralized framework.

Finally, the client minimizes a total loss concerning θ only, following the exploitation in Equation (8), excluding the FPG component. The loss is given by

$$\mathcal{L}_{total} = \mathcal{L}_t + \lambda_f \mathcal{L}_f, \tag{10}$$

where $\mathcal{L}_t$ refers to a balanced risk minimization (BRM) loss as BSM (Ren et al., 2020), and λ_f as the free weighting factor.

3.2.2. *Federated Long-Tailed Study*

In FL settings, estimating both the global long-tailed and intra-class distribution imbalance can be difficult due to the decentralization of data. Prior studies on long-tailed recognition that rely on identifying head or tail classes through prediction confidences or classifier weights (Kang et al., 2020; Kobayashi, 2021) can impede representation learning by exacerbating intra-class imbalance (Ju et al., 2022; Tang et al., 2022). In Fed-LT, both extreme class imbalances and inter-client intra-class variations can lead to client drift. For instance, a class attribute imbalance may surface across clients due to differences in findings, scanners, or populations. As a result, estimating the global LT distribution in medical images within the FL framework is a challenge that is yet to be explored.

We study the normalized distribution of $\mathcal{L}_f^{\theta}$ (Equa-

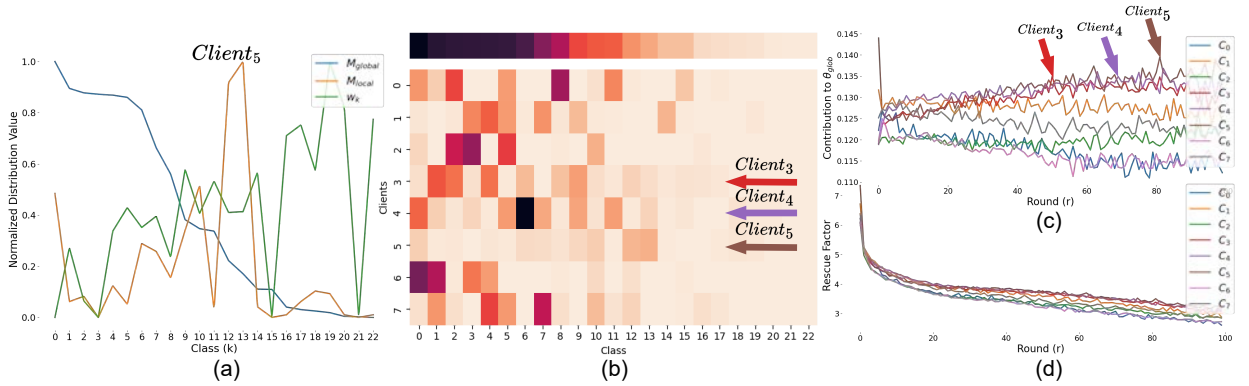


Figure 6: Analysis of DLMA: (a) The globally aggregated class counts (M_{global}), and client five local class counts (M_{local}) with w_k in one round. (b) Client's Data Distribution, (c) Client's Contribution to θ_{glob} throughout rounds, (d) Rescue Factor (RF) on different clients throughout rounds

tion (7)) in each class k with M_k total samples as:

$$\mathcal{L}_k^{\theta} = \frac{1}{M_k} \sum_{i=1}^{M_k} L_f(x_{k,i}). \tag{11}$$

At the beginning of each round in the federated learning process, each client receives the model from the global server θ_{global}, which is evaluated on the local data in each client, generating a loss value w_k, where $w_k = \mathcal{L}_k^{\theta_{glob}}$. The factor w_k can help capture the distribution difference between the global server and the free model on each client's local data. This divergence can provide insights into the sensitivity of the global model, θ_{glob}, in effectively capturing the specific classes in each client's local data during federated learning. A high w_k indicates the failure of θ_{glob} in capturing a local client class k. In Fed-LT, we can see that w_k is inversely proportional to the global LT distribution, even if the local client is not necessarily LT (See Figure 6 (a)).

A client updates its local model, θ' with the received global model, θ_{glob} and takes subsequent optimization steps for E local epochs while estimating $\hat{w}_k$, where $\hat{w}_k = \sum_{e=1}^{E} \mathcal{L}_k^{\theta'}$. The factor $\hat{w}_k$ can help to capture how well the information from m_ξ has been distilled to each of the local client's classes (distillation belief).

We can then correct this estimation, $\hat{w}_k$, with a global observation w_k to generate a rescue factor, RF, at each client in every round.

$$RF = \sum_{k=1}^{K} w_k \hat{w}_k. \tag{12}$$

A higher RF indicates that the client has information that the global model has not appropriately captured.

3.2.3. Dynamic LT-Aware Model Aggregation (DLMA)

Inspired by the fact that client-specific models should contribute more to the global server to capture local variance, we propose a novel dynamic LT-aware model aggregation (DLMA). We use our proposed RF to indicate client-specific models that should contribute more to the global model than client-generic models to capture their class variations (Client 5 in brown in Figure 6 (b) have mostly tail classes and contribute the most to θ_{glob} in Figure 6 (c)). While our proposed RF can be used for biased client selection (Jee Cho et al., 2022), we use it to aggregate a global model. Instead of aggregating based on the weighted samples as in FedAvg (McMahan et al., 2017), we propose to weight the global model, θ_{glob}, based on the RF value as follows:

$$\bar{RF}_c = \frac{RF_c}{\sum_j RF_j}, \text{ and } \theta_{glob}^{r+1} = \sum_{c=1}^{C} \bar{RF}_c \theta'_c. \tag{13}$$

We show in Figure 6 (d) that the rescue factor for all clients is decreasing throughout rounds. This highlights the ability of DLMA to accommodate different clients. Also, the *scalar*, RF, does not reveal the input data distribution. (See Algorithm 1).

4. Experiments

4.1. *FoPro-KD*

4.1.1. Datasets

ISIC-LT is a challenging long-tailed skin lesion classification dermatology dataset from ISIC (Combalia et al., 2019). The original ISIC dataset consists of eight classes, and we create a long-tailed version of it following (Ju et al., 2022) using a Pareto distribution sampling approach. To ensure class imbalance and rare disease diagnosis, we set the class imbalance ratio between the majority class and the minority class in training set to be 100, 200, 500 and select 50 and 100 images from each class for the validation and test sets, respectively, from the remaining samples. We assess the model performance on the held-out test set. Results for each method are averaged over five runs, each with a different sampled train, validation, and held-out test set. To assess the

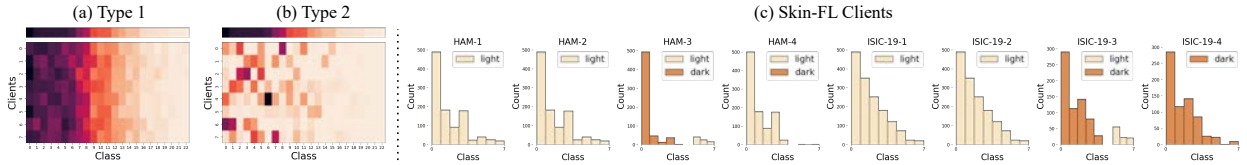


Figure 7: Decenteralized Data Division across HyperKvasir with (a) Type-1 and (b) Type-2, and (c) the attribute setting of ISIC-LT

Algorithm 1 Pseudocode for FedFree.

Notations total number of clients (C), server (S), total communication rounds (R), local epochs (E), learning rate (η), and a set of client's data sliced into batches of size B ($\mathcal{B}$).
ServerExecution:
Init θ^1_{glob}
for *each round* $r = 1, ..., R$ **do**
  for *client* $c \in C$ *in parallel* **do**
    $\theta_c, RF_c \leftarrow$**LocalUpdate**$(\theta^r_{glob})$;
  end for
  $\theta^{r+1}_{glob} \leftarrow$**DLMA**$(RF_c, \theta'_c$,c = 1 to C); // Equation (13)
end for
Return θ^R_{glob}
LocalUpdate (θ_{glob}):
Init $\hat{w}_k = 0$;
Init $w_k = \mathcal{L}_k^{\theta_{glob}}$;
for *each local epoch* $e = 1, ..., E$ **do**
  for *each batch* $b \in \mathcal{B}$ **do**
    $\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda_f \mathcal{L}_f$; // Equation (10)
    $\theta' \leftarrow \theta' - \eta \nabla \mathcal{L}_{total}$;
    $\hat{w}_k \leftarrow \hat{w}_k + \mathcal{L}_f(b_k)$; // running free loss mean for each class k
  end for
end for
$RF = \sum_{k=1}^{K} w_k \hat{w}_k$; // RF $\uparrow \approx$ divergence $\theta_{glob}, m_\xi \uparrow$
Return θ', RF

model performance on the balanced test set, we follow recent guide (Reinke et al., 2022) to report the Mathew-correlation coefficient ("MCC"), accuracy ("Acc"), and f1-score.

Hyperkvasir is a long-tailed dataset of 10,662 gastrointestinal tract images comprising 23 classes representing different anatomical and pathological landmarks and findings. To analyze the long-tailed distribution, we categorize the 23 classes into three groups: Head (with over 700 images per class), Medium (with 70 to 700 images per class), and Tail (with fewer than 70 images per class) based on their class counts. Notably, the Tail class includes a distinct class for Barrett's esophagus, which presents a short segment and is considered a pre-malignant condition that may progress to cancer. Additionally, the Tail classes encompass two transitional grades of ulcerative colitis, an inflammatory bowel disease, and the terminal ileum, which confirms a complete colonoscopy but cannot be differentiated endoscopically from parts of the small bowel. Since the official test set only contains 12 classes, we follow the evaluation approach of BalMixUp (Galdran et al., 2021) and assess our model's performance using a stratified 5-fold cross-validation method. To assess the performance with a high imbalance test set, We report the balanced accuracy ("B-Acc") that considers the average per class accuracy and denotes the performance of the few-shot division ("Head', "Medium", "Tail") and their average results denoted as "All".

4.1.2. Implementation Details

For both datasets, we use checkpoints pre-trained on MoCo-RN50 (He et al., 2020) available online as the free lunch models trained on ImageNet for compressing its generalization capabilities unless otherwise stated. We use Adam optimizer with a learning rate of $3e-4$ for all methods on the ISIC-LT dataset. On the other hand, we follow (Galdran et al., 2021) for the HyperKvasir dataset and use SGD with a cosine annealing scheduler (Loshchilov and Hutter, 2017) with a maximum learning rate of 0.01. For both datasets and all methods, we use a ResNet-18 as the target model with a batch size of 32 and apply augmentations techniques such as random crop and flipping. Images are resized to 224x224, and we train all methods until there is no further increase in the validation set for 20 epochs with a total of 100 epochs. To ensure a fair comparison between different methods, we keep all hyperparameters the same. We set λ_f to 3, μ to 10, and γ to 0.3 on both datasets. For every five training epochs exploited, we explore the pre-trained model for one epoch to balance the training process.

4.1.3. Baselines

Our experimental evaluation compares the performance of our proposed FoPro-KD method against several state-of-the-art long-tailed learning approaches. Specifically, we evaluate (1) re-sampling (RS) and re-weighting (RW) techniques, (2) various data augmentation techniques including APR (Chen et al., 2021b), MixUp (Zhang et al., 2018), and its balanced version (BalMixUp) (Galdran et al., 2021), specifically designed for medical image classification. (3) Modified Loss re-weighting schemes including Class balancing (CB) loss (Cui et al., 2019), and label-distribution-aware margin (LDAM) loss with curriculum delayed

Table 1: Experimental results on long-tailed skin lesion classification (ISIC-LT) with different class imbalance ratios.

Method	Class Imbalance Ratio								
	1:100			1:200			1:500		
	MCC	Acc	F1-Score	MCC	Acc	F1-Score	MCC	Acc	F1-Score
CE	57.64 (±1.6)	62.15 (±1.4)	65.52 (±1.4)	53.71 (±1.7)	58.33 (±1.5)	62.72 (±1.2)	44.9 (±2.2)	50.22 (±1.9)	55.83 (±2.0)
RS	59.46 (±1.0)	63.9 (±0.9)	67.04 (±0.6)	55.53 (±1.6)	60.35 (±1.5)	63.71 (±1.4)	48.54 (±1.4)	53.73 (±1.2)	59.15 (±1.1)
RW	56.03 (±2.3)	61.2 (±1.9)	63.17 (±2.2)	52.22 (±1.6)	57.95 (±1.4)	59.48 (±1.5)	46.77 (±0.4)	52.8 (±0.4)	55.36 (±0.7)
CE-IFL (Tang et al., 2022)	60.58 (±1.8)	64.72 (±1.6)	67.96 (±1.6)	57.06 (±2.2)	61.45 (±1.9)	65.4 (±1.7)	47.26 (±1.6)	52.32 (±1.5)	57.99 (±1.7)
CE-EKD (ours)	61.37 (±1.8)	65.42 (±1.6)	68.49 (±1.5)	57.57 (±1.1)	61.9 (±0.9)	65.41 (±1.2)	49.16 (±1.9)	54.2 (±1.8)	59.24 (±1.4)
CB (Cui et al., 2019)	57.28 (±2.3)	62.23 (±2.1)	64.36 (±1.6)	53.58 (±2.1)	58.9 (±2.1)	61.27 (±1.6)	47.16 (±1.2)	53.17 (±1.1)	55.9 (±1.6)
LDAM-DRW (Cao et al., 2019)	60.27 (±0.7)	64.88 (±0.6)	66.17 (±0.7)	55.85 (±1.6)	60.98 (±1.5)	62.25 (±1.3)	50.34 (±1.1)	55.98 (±0.8)	57.95 (±1.3)
BSM (Ren et al., 2020)	63.88 (±1.9)	68.15 (±1.7)	69.25 (±1.6)	60.47 (±1.6)	65.12 (±1.4)	66.2 (±1.2)	53.61 (±1.1)	59.02 (±1.0)	60.27 (±0.9)
MixUp (Zhang et al., 2018)	55.53 (±1.8)	59.91 (±1.9)	64.33 (±1.0)	48.96 (±2.1)	53.59 (±2.2)	59.68 (±1.5)	43.03 (±1.6)	48.12 (±1.5)	54.36 (±1.1)
APR (Chen et al., 2021b)	57.05 (±1.5)	61.65 (±1.4)	65.23 (±1.0)	52.84 (±0.9)	57.67 (±0.9)	61.64 (±0.9)	45.5 (±1.2)	50.78 (±1.1)	56.5 (±1.1)
BalMixup (Galdran et al., 2021)	61.35 (±1.8)	65.5 (±1.5)	68.46 (±1.5)	56.36 (±3.9)	61.0 (±3.5)	64.37 (±3.5)	50.26 (±1.1)	55.3 (±1.1)	60.29 (±0.7)
BSM-APR (Chen et al., 2021b)	63.29 (±2.8)	67.7 (±2.5)	68.59 (±2.4)	61.07 (±2.0)	65.7 (±1.8)	66.64 (±1.6)	52.94 (±1.9)	58.48 (±1.6)	59.93 (±2.0)
BSM-IFL (Tang et al., 2022)	65.01 (±1.9)	69.05 (±1.7)	70.48 (±1.5)	60.42 (±2.3)	64.95 (±2.0)	66.6 (±1.7)	54.12 (±1.8)	59.15 (±1.5)	61.69 (±1.9)
BKD (Zhang et al., 2023)	62.24 (±1.6)	66.55 (±1.6)	68.35 (±0.9)	63.06 (±1.4)	67.42 (±1.2)	68.32 (±1.3)	54.25 (±1.3)	59.59 (±1.1)	60.5 (±1.2)
FoPro-KD (ours)	**68.33 (±2.3)**	**71.8 (±2.0)**	**73.88 (±1.9)**	**66.08 (±1.5)**	**69.8 (±1.3)**	**71.91 (±1.2)**	**57.33 (±1.5)**	**61.9 (±1.5)**	**64.43 (±1.3)**

reweighting (DRW) (Cao et al., 2019), and the balanced softmax (BSM) (Ren et al., 2020) (4) A recent curriculum-based method, balanced Knowledge Distillation (BKD) (Zhang et al., 2023).

*4.2. **FedFree***

4.2.1. Dataset

HyperKvasir-FL We follow the same splitting strategy as proposed by (Shang et al., 2022) to split the HyperKvasir to eight clients, resulting in Type 1 and Type 2 as shown in Figure 7 (a) and (b) respectively. Type 1 indicates that clients are identically distributed (iid) following a long-tail distribution that matches the global distribution. However, Type 2 indicates that split data clients are non-identically distributed (non-iid) relative to their class count following a Dirichlet distribution. We use Dirichlet distribution with $\alpha = 0.5$ to simulate Type 2. We assess all methods' performance using stratified 5-fold cross-validation.

ISIC-LT Attribute Split To investigate the inter-client intra-class variations within a specific attribute, we propose conducting our study on the ISIC-LT dataset using the skin color attribute. This choice aims to replicate real-world attributes and their distributions. we leverage the publicly available skin tone labeling provided by Bevan and Atapour-Abarghouei (2022), after removing any duplicate samples. Then, we divide the dataset into two distributions, namely HAM-1000 (Tschandl et al., 2018) and ISIC (Combalia et al., 2019), to simulate heterogeneity among clients. We further categorize the clients based on two attributes: light and dark skin tones. The division of the 8 clients is determined by the attribute, class count, and dataset distribution as depicted in Figure 7 (c). For instance, client 1-4 is derived from the HAM-1000 distribution, with client HAM-3 characterized by a dark skin tone for classes 1-5 ("Head") and a light skin tone for classes 6-8 ("Tail"). On the other hand, clients 4-8 are obtained from the ISIC-19 distribution, with client 8 encompassing a dark "Head" and "Tail". Additionally, we split the data between each client for training, validation, and testing with 70%, 15%, and 15%, respectively. The global validation and global test set is an aggregation of the local validation and test. To evaluate the performance of attributes, we report the "B-Acc" separately for each attribute ("Light", "Dark") within each distribution ("HAM-1000", "ISIC-19"), and the average of these scores "Avg". Additionally, we report the overall "B-Acc" across all attributes and distributions.

4.2.2. Implementation Details

To simulate the FL setting, we adopt a torch multiprocessing strategy and deploy each local client on an NVIDIA RTX-3090 with each client having the same implementation details as in Section 4.1.2. Finally, for Hyperkvasir and the ISIC-FL, we train eight clients for 200 and 100 communication rounds respectively, or until the global model convergence.

4.3. Baselines

We compare our methods with FL methods. Specifically, we evaluate FedAvg (McMahan et al., 2017), FedProx (Li et al., 2020), MOON (Li et al., 2021). We also integrate Focal (Lin et al., 2017), LDAM-DRW (Cao et al., 2019), and BSM (Ren et al., 2020) into the FedAvg framework and rename them as Focal-FL, LDAM-FL, and BSM-FL respectively. Additionally, we compare our results with a label-distribution skew FL method, FedLC (Zhang et al., 2022), and a federated-LT method CReFF (Shang et al., 2022).

5. Results

*5.1. **FoPro-KD***

5.1.1. Performance on ISIC-LT

We present the performance of our proposed FoPro-KD approach for long-tailed skin lesion classification on the ISIC-LT dataset in Table 1. Our approach outperforms all baselines across all class imbalance ratios and evaluation metrics, demonstrating its effectiveness. FoPro-KD improves the performance of the naive cross

entropy by 10.7%, 12.4%, and 12.4% on the "MCC" over the balanced test set for class imbalance ratios of 1:100, 1:200, and 1:500, respectively. Compared to the baseline, BSM (Ren et al., 2020), FoPro-KD improves the "MCC" being sensitive for class imbalance by 4.5%, 5.6%, and 3.7% for imbalance ratios of 1:100, 1:200, and 1:500, respectively. Furthermore, it increases the performance of the baseline, BSM (Ren et al., 2020), by 3.7%, 4.7%, and 2.9% on the "Acc" metric for class imbalance ratios of 1:100, 1:200, and 1:500, respectively. Compared to the best-performing baseline on imbalance ratios 1:200 and 1:500, BKD (Zhang et al., 2023), our method outperforms it by 6.0%, 3.0%, and 3.0% on the "MCC" for the three imbalance ratios, respectively. Notably, our approach outperforms BKD on the f1-score with 3.6% and 3.9% performance gains over the imbalance ratios 1:200 and 1:500 without additional pre-training on the target dataset.

It is worth mentioning that our proposed EKD used with the naive cross-entropy loss improves performance by 3.7%, 3.9%, and 4.3% on the "MCC" metric for class imbalance ratios of 1:100, 1:200, and 1:500, respectively, without the need for FPG or special loss re-weighting or re-sampling, demonstrating the need to leverage the free lunch models for the long-tail problems in an effective way.

Table 2: Experimental results on long-tailed Gastrointestinal image recognition. The top-1 accuracy is reported using a shot-based division ("Head", "Medium", "Tail") to address test set imbalance, and their average "All", along with the resilient metric "B-Acc" for class imbalance.

Method	Metrics				
	Head	Medium	Tail	All	B-Acc
CE	93.14 (±0.7)	74.7 (±1.2)	4.05 (±4.8)	57.3 (±1.3)	58.81 (±1.1)
RS	88.89 (±3.9)	72.37 (±3.2)	11.38 (±10.4)	57.55 (±1.8)	58.84 (±1.6)
RW	87.43 (±1.8)	70.04 (±2.5)	20.28 (±7.6)	59.25 (±2.0)	60.19 (±1.8)
CB (Cui et al., 2019)	88.22 (±1.5)	70.36 (±1.7)	18.04 (±9.8)	58.88 (±2.7)	59.88 (±2.5)
LDAM-DRW (Cao et al., 2019)	92.53 (±0.6)	69.4 (±1.5)	24.55 (±9.1)	62.16 (±2.5)	62.79 (±2.2)
BSM (Ren et al., 2020)	91.4 (±0.7)	65.96 (±3.0)	26.54 (±7.7)	61.3 (±1.9)	61.7 (±1.6)
MixUp (Zhang et al., 2018)	94.23 (±0.6)	75.08 (±1.2)	3.93 (±3.3)	57.75 (±1.0)	59.25 (±0.9)
BalMixUp (Galdran et al., 2021)	92.16 (±1.1)	74.57 (±1.7)	8.44 (±3.8)	58.39 (±1.1)	59.8 (±0.9)
BKD (Zhang et al., 2023)	92.53 (±0.9)	69.88 (±5.0)	17.43 (±12.6)	59.95 (±2.7)	60.81 (±2.3)
FoPro-KD (ours)	92.78 (±2.0)	68.08 (±6.5)	31.9 (±8.5)	**64.25 (±0.8)**	**64.59 (±0.9)**

5.1.2. Performance on HyperKvasir

We present the experimental results of our method on the long-tailed gastrointestinal image recognition in Table 2. Our approach outperformed the naive cross-entropy method by 7.0% and 5.8% for the highly imbalanced test-set and increased the performance of the baseline (Ren et al., 2020) by 2.9% and 2.9% on the "All" and "B-Acc" metrics respectively. Moreover, our method achieved the highest performance on the "Tail" (31.9%), highlighting its ability to capture rare diseases.

Our method outperformed the state-of-the-art BKD (Zhang et al., 2023) on the HyperKvasir dataset. BKD relies on distilling a pre-trained teacher model over the target dataset, which can amplify bias over the head classes if the teacher model fails to capture the tail classes. In contrast, our approach leverages the discriminative generalizable features of free lunch models. Specifically, our approach outperformed BKD by 4.3% and 3.8% over "All" and "B-acc", respectively.

5.1.3. Ablation

Effectiveness of EKD and FPG In Table 3, we present an ablation study of our proposed components over the ISIC-LT. Our approach combines a Fourier prompt generator (FPG) with effective knowledge distillation (EKD) to exploit the pre-trained model. Our experimental results on the ISIC-2019 dataset demonstrate that EKD alone improves performance by 1.5%, 3.6%, and 2.2% on the "MCC" for the three imbalance ratios, respectively. By adding FPG, we achieve even higher performance gains of 4.5%, 5.6%, and 3.7% on the "MCC" for class imbalance ratios of 1:100, 1:200, and 1:500 compared to the baseline, BSM (Ren et al., 2020).

Table 3: Ablation of FLKD and FPG on three imbalance ratios on ISIC-LT

	EKD	FPG	Metric		
			MCC	Acc	F1-Score
			ISIC-LT (1:100)		
BSM (Ren et al., 2020)	×	×	63.88 (±1.9)	68.15 (±1.7)	69.25 (±1.6)
w/ **EKD (ours)**	✓	×	65.36 (±3.3)	69.47 (±2.9)	70.42 (±2.9)
FoPro-KD (Ours)	✓	✓	**68.33 (±2.3)**	**71.8 (±2.0)**	**73.88 (±1.9)**
			ISIC-LT (1:200)		
BSM (Ren et al., 2020)	×	×	60.47 (±1.6)	65.12 (±1.4)	66.2 (±1.2)
w/ **EKD (ours)**	✓	×	64.08 (±1.4)	68.35 (±1.2)	69.19 (±1.3)
FoPro-KD (Ours)	✓	✓	**66.08 (±1.5)**	**69.8 (±1.3)**	**71.91 (±1.2)**
			ISIC-LT (1:500)		
BSM (Ren et al., 2020)	×	×	53.61 (±1.1)	59.02 (±1.0)	60.27 (±0.9)
w/ **EKD (ours)**	✓	×	55.81 (±1.6)	60.92 (±1.4)	62.02 (±1.3)
FoPro-KD (Ours)	✓	✓	**57.33 (±1.5)**	**61.9 (±1.5)**	**64.43 (±1.3)**

Our proposed EKD and FPG methods provide complementary benefits for improving the performance of the target model in the long-tailed setting. While FPG helps to explore the pre-trained model's latent space by explicitly asking what frequency patterns it wants in the input, EKD helps to exploit the pre-trained model's generalizable representation. By leveraging pre-trained models' frequency patterns, our approach achieves the best performance on the ISIC-LT dataset and HyperKvasir dataset, highlighting the importance of utilizing pre-trained models for medical image classification with long-tailed class distributions.

Ablation of Free Factor We present an ablation study of the weighting factor, λ_f, for the exploitation proposed in Equation (8), with experiments conducted on the ISIC-LT imbalance factor 1:500 without our proposed FPG. The results are summarized in Table 4.

Table 4: Exploitation λ_f ablation without FPG on the ISIC-LT (Acc)

Method	ISIC-LT (1:500)			
	$\lambda_f = 0$	$\lambda_f = 1$	$\lambda_f = 3$	$\lambda_f = 5$
EKD	59.02 (±1.0)	59.52 (±2.4)	**60.92 (±1.4)**	60.47 (±1.6)

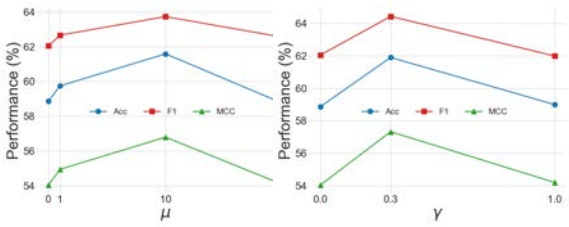
We find that using effective knowledge distillation (EKD) with a factor of $\lambda_f = 3$ improves the performance on the ISIC-LT dataset compared to the baseline ($\lambda_f = 0$), (Ren et al., 2020), achieving an "Acc" gain of

1.9%. However, a higher value of λ_f can deviate from the learning objective.

Table 5: Ablation of Exploration on the ISIC-LT 1:500 dataset

	$\mathcal{L}_{inv}$	$\mathcal{L}_{adv}$	Metric		
			MCC	Acc	F1
EKD (ours)	×	×	55.81 (±1.6)	60.92 (±1.4)	62.02 (±1.3)
Explore only	✓	×	56.80 (±1.4)	61.59 (±1.2)	63.73 (±1.8)
FoPro-KD	✓	✓	**57.33 (±1.5)**	**61.9 (±1.5)**	**64.43 (±1.3)**

Effectiveness of FPG To evaluate the importance of $\mathcal{L}_{inv}$, we perform an ablation study and report our results in Table 5. We find that learning the FPG and exploring the frozen pre-trained model with only $\mathcal{L}_{inv}$ leads to an improvement over our proposed EKD with an increase of 1.0 % and 1.7% on "MCC" and f1-score, respectively. Moreover, when using iterative adversarial knowledge distillation (AKD) along with $\mathcal{L}_{inv}$, we achieve the best performance with a notable gain of 1.5%, 1.0%, 2.4% on the "MCC", "Acc", and F1 score respectively, compared to our proposed EKD. While $\mathcal{L}_{inv}$ ensures that the synthesizable Fourier amplitudes are representative of what the free lunch model wants, capturing the frequency patterns in the frequency bands it was trained on, $\mathcal{L}_{adv}$ is responsible for further exploring the latent space of the frozen model and making the frequency prompts more diverse than the ones previously distilled to the target model. (See 5.1.3 for FPG output ablation)



(a) (b)

Figure 8: Sensitivity to μ and γ on the ISIC-LT Imbalance Ratio 1:500

Sensitivity of Balancing Regularization Batch normalization (BN) statistics are necessary for learning the Fourier prompts (FPG) in our proposed method. Similar to deep inversion and data-free knowledge distillation approaches (Fang et al., 2021), without BN, the FPG can be limited to balancing regularization. we perform ablation experiments on the balancing regularization weighting factor μ for the exploration phase proposed in Equation (6) over the extremest ISIC-LT setting (1:500). As shown in Figure 8 (a), we observe that a value of $\mu = 10$ increases the performance by 2.6%, 2.4%, and 2.5% on the "MCC", "Acc", and F1, respectively. Without using μ, the exploration phase is limited to the BN statistics without activation of the free-lunch model latent space, which can limit the representation transfer. A high value of μ, however, can negatively impact performance by encouraging the network to output a uniform distribution that is not discriminative nor informative.

Sensitivity of AKD Next, we investigate the effect of the adversarial factor γ proposed in Equation (9) on the performance of the extremest ISIC-LT setting (1:500). We found that a low value of γ (e.g., $\gamma = 0.3$) can enhance performance by making the Fourier prompts more diverse with iterative adversarial training, increasing the performance by 1.0% 2.0% on the F1-score and "MCC", as shown in Figure 8 (b). On the other side, a high value of γ (e.g., $\gamma = 1$) results in a 2.0% drop in the F1-score. It is worth noting that, unlike other adversarial training approaches in domain adaptation, our focus is not on adversarial training but on synthesizing images based on the frozen pre-trained model by our proposed FPG. A lower value of γ ensures the diversity of generated prompts, whereas a higher value may result in FPG generating worst-case images with random amplitudes that the frozen pre-trained model cannot comprehend, leading to a decrease in overall performance.

FPG is conditional on both the input and the frozen free-lunch models

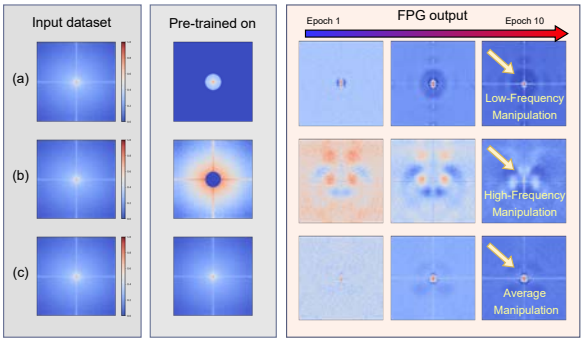


Figure 9: Average FPG generated prompts in three scenarios of pre-training f on different frequency components of the ISIC-LT dataset. (a) Pre-training f on only the low-frequency components. (b) Pre-training f on only the high-frequency components. (c) Pre-training f on all-frequency components.

In Figure 9, we demonstrate the behavior of our FPG with different settings. In (a), we trained the frozen pre-trained model, f, on only the low-frequency components of the ISIC-LT dataset. We observed that the FPG converged to a similar average amplitude as the input dataset but with different surpassing and amplification in the low-frequency parts that are conditional on f. (b) shows the FPG's behavior when f was trained on only the high-frequency components of the ISIC-LT dataset. We found that the FPG attends to the different frequencies in their higher frequencies that f has captured. Finally, in (c), we trained f on all frequency components of the input dataset. Interestingly, we found that the average amplitude generated by the FPG does not fully reduce to the amplitude of the source dataset, although it is conditional on the input dataset. This is because we do not have any prior knowledge of what fre-

quency patterns in what frequency bands the pre-trained model extracts from the dataset in the pre-training stage. Nonetheless, FPG was able to amplify or suppress certain frequencies to provide understanding and interpretation of the behavior of pre-training models.

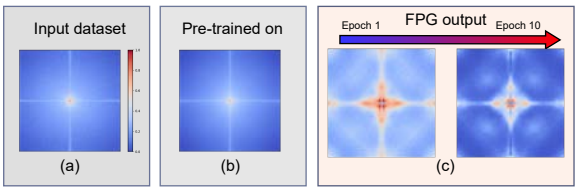


Figure 10: (a) Average ISIC-19 Fourier amplitude spectrum. (b) Average ImageNet Fourier amplitude spectrum. (c) FPG output in different epochs.

We show in Figure 10 when the frozen pre-trained model f is pre-trained on the ImageNet dataset with the average ImageNet Fourier spectrum. The generated output of our FPG is conditioned on both the frequency patterns extracted by the pre-trained model f from the ImageNet Fourier spectrum and the frequency patterns of the target input dataset.

Our proposed Fourier Prompts Generator (FPG) is designed for understanding and interpreting the behavior of pre-trained models. Unlike prior methods that rely on adding noise to synthesize worst-case images, the FPG is conditional on both the input dataset and the frozen pre-trained model. We demonstrate in our analysis that the FPG can be used with only the L_{inv} and no adversarial training. Our method leverages the different frequency patterns captured in the pre-training stage of the network to amplify or suppress certain frequencies. By exploring these patterns with the FPG, we can gain insights into the specific input preferences of the pre-trained model that enable better representational transfer and interoperability. However, to enhance the generalizability and interpretability of our approach, further investigation is required into the distinct patterns of frequencies captured by the pre-trained model f, and their differences on more sophisticated models.

EKD benefits LT even with smaller models The learning of the target model can be limited with an upper bound to the capacity of the free lunch model, f, and the MLP projector, and the information gained from f to the target task. However, we demonstrate in Table 6 that such limitations do not adversely affect the performance of the minority class on ISIC-19 LT ("Tail"), with linear probed (LP) supervised ImageNet weights achieving 41.89% and EKD achieving 55.93% on the "Tail" accuracy.

Our experiments presented in Table 6 demonstrate that our proposed EKD method can improve the performance of the target task even with smaller models. Specifically, we show that when given a target model g and its pre-trained version as the free lunch model f, EKD can benefit the tail classes using the frozen features from f despite f having the same capacity as g and being pre-trained on ImageNet. We observed a performance gain of 3.8% on tail class accuracy and 0.94% on "MCC" compared to the best-performing baseline initialized with ImageNet weights. It is worth mentioning that these results are averaged over 5 runs over the 3 class imbalance ratios (15 experiments).

This phenomenon arises because fine-tuning can distort the pre-trained features, leading to a drop in generalization performance. However, the target model can further enhance its performance by using free discriminative distribution during training. While EKD can improve the performance of the target task even with smaller models, the best performance is achieved when using ResNet50 (RN50) as the free lunch model (f), with a performance gain of 1.85% and 1.81% on "MCC" compared to the baseline, BSM (Ren et al., 2020), when the target model is initialized randomly (None) or with ImageNet weights respectively.

While most empirical evaluations ignore pre-trained initialization to provide fair and better convergence analysis, initialization unsurprisingly increases the averaged performance by 6.92% and 6.96% on "MCC" for the baseline and our proposed EKD, respectively.

Table 6: Effective Knowledge Distillation (EKD) with varying free lunch models. Results are averaged across 5 runs and across the three imbalance ratios (1:100, 1:200, 1:500) on the ISIC-LT dataset.

Setting					Metric (%)	
Method	Target	Target Init	Free Lunch	Free Lunch Init	Tail	MCC
LP	None	None	RN-50	Sup-ImageNet	41.89	48.77
LP	None	None	RN-50	MoCov2	48.72	49.85
BSM	RN-18	None	None	None	46.6	59.32
EKD	RN-18	None	RN-50	Sup-ImageNet	51.47	61.13
BSM	RN-18	ImageNet	None	None	52.07	66.24
EKD	RN-18	ImageNet	**RN-50**	Sup-ImageNet	**55.93**	**68.09**
EKD	RN-18	ImageNet	RN-18	Sup-ImageNet	55.87	67.18

Free Lunch model Ablation We present an ablation study of full fine-tuning and linear probing of the free lunch models pre-trained solely on ImageNet without any knowledge from our target task, used with both the naive cross-entropy (CE) and the baseline, BSM (Ren et al., 2020).

As shown in Table 7, our experiments on linear probing models reveal that linear probing models initialized with MoCoV2 outperform models initialized with supervised ImageNet pre-training, with a performance gain of 2.26% and 6.83% on "Head" and "Tail" accuracy, respectively. This is because MoCoV2 learns competitive generalizable discriminative representations that benefit the tail classes in linear probing. Thus, our approach of leveraging models pre-trained solely on natural images without any prior knowledge from the target dataset proves to be effective.

On the other hand, full fine-tuning with the naive-cross entropy benefits the head classes more than linear probing with a gain of 25.18% on the "Head" but comes with a performance drop of 12.7% on the "Tail" in com-

parison with the best-performing linear probing. These models are computationally heavy for full-fine tuning and deployment, having almost two times the number of parameters as the target model (23 million vs 11 million parameters). Therefore, we propose EKD distilling and compressing such generalization capabilities to smaller models.

Table 7: Linear Probing (LP) and Fine-Tuning (FT) Accuracy of free lunch models with naive cross-entropy (CE), averaged over 5 runs across the three imbalance factors for the ISIC-LT dataset. The table includes the accuracy of the majority class ("Head") and the accuracy of the minority class ("Tail")

Setting		Metric (%)			
Method	Free Lunch	Head	Tail	MCC	Acc
LP	Sup-ImageNet	65.36	41.89	48.77	54.96
LP	MoCov2	67.62	**48.72**	49.85	55.99
FT	Random	88.23	22.88	49.46	54.63
FT	MoCov2	89.93	27.6	56.0	60.48
FT	Sup-ImageNet	**92.8**	35.93	**62.51**	**66.34**

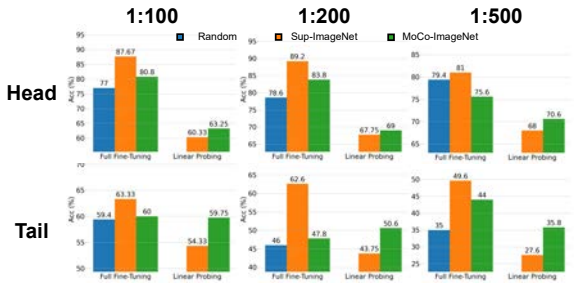
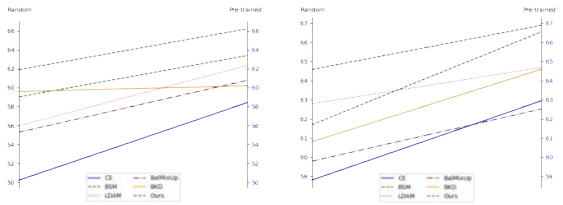


Figure 11: Bar plots illustrating the Linear Probing and Fine-Tuning Accuracy of free lunch models with baseline (BSM) (Ren et al., 2020) across three imbalance factors (1:100, 1:200, and 1:500) for the ISIC-LT dataset.

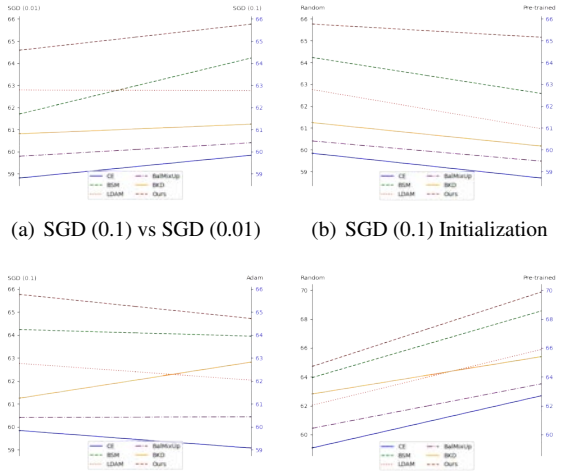
In Figure 11, we show the results of full fine-tuning and linear probing of the free lunch models utilized with the best-performing baseline, BSM (Ren et al., 2020). We observed that the linear probing of the free lunch model (RN50) trained with contrastive learning (MoCov2) works better than the supervised version linear probed. The performance gain is 2.9%, 1.2%, and 2.6% on the head class and 5.4%, 6.9%, and 8.2% on the tail class for the three imbalance factors, respectively. This gain is attributed to the generalizable features from MoCo-v2 as reported previously. However, when fully fine-tuning the weights of the supervised ImageNet with the baseline (Ren et al., 2020), we observed compelling performance over both the head and tail classes compared to fully fine-tuning initialized with MoCov2. The performance gain is 6.8%, 5.4%, and 5.4% on the "Head" and 6.9%, 14.8%, and 5.6% on the "Tail" for the three imbalance factors, respectively. This gain can be attributed to the supervised ImageNet's weights being already suited for supervision signals. These results are consistent for all three imbalance factors.

Our experiments have shown unsurprisingly that initialization methods play a crucial role in the performance of deep learning models on imbalanced datasets. While we show empirically that linear probing with contrastive learning approaches works best due to their superior generalization capabilities, we find that the supervised ImageNet initialization provided the best initialization performance for full fine-tuning of the free lunch model.



(a) ISIC-LT (1:500) (b) HyperKvasir

Figure 12: Initialization of target models with ImageNet pre-trained weights and random weights for all methods on ISIC-LT and HyperKVasir datasets. The left y-axis represents the performance of models initialized with random weights, while the right y-axis represents the performance of models initialized with ImageNet pre-trained weights.



(a) SGD (0.1) vs SGD (0.01) (b) SGD (0.1) Initialization

(c) SGD (0.1) vs Adam (d) Adam Initialization

Figure 13: Comparison of different optimizers and different initialization on the HyperKvasir dataset. (a) and (c) depict experiments with different optimizers, while (b) and (d) represent experiments with different initialization.

Effectiveness of weight initialization In Figure 12, we demonstrate the importance of weight initialization when dealing with long-tail learning. Recent works have mostly ignored such initialization steps in their empirical experiments to provide fair and better convergence analysis. We agree that pre-trained model weights are architecture-dependent and may not be available for smaller models with the rising of large publicly available pre-trained models. However, if the smaller pre-trained model's weights are available, it

Table 8: Comparison with other methods on HyperKvasir Dataset. We comprehensively evaluate different LT methods in FL. All clients are initialized with ImageNet pre-trained weights; each result is averaged over five runs.

	Type 1					Type 2				
Methods	Head	Medium	Tail	**All**	**B-acc**	Head	Medium	Tail	**All**	**B-acc**
Federated Learning Methods (FL-Methods)										
FedAvg (McMahan et al., 2017)	94.1 ± 1.3	72.9 ± 1.3	3.1 ± 0.9	56.69 ± 0.6	58.1 ± 0.6	86.2 ± 2.7	70.3 ± 0.5	8.0 ± 1.2	54.83 ± 1.0	56.17 ± 0.9
FedProx (Li et al., 2020)	94.6 ± 0.4	72.1 ± 0.2	3.0 ± 1.2	56.58 ± 0.4	57.93 ± 0.4	88.1 ± 2.2	73.1 ± 2.7	3.6 ± 2.5	54.93 ± 1.3	56.51 ± 1.3
MOON (Li et al., 2021)	94.7 ± 0.7	74.6 ± 0.4	4.0 ± 1.8	57.77 ± 0.6	59.23 ± 0.6	84.4 ± 3.6	73.1 ± 1.6	5.5 ± 2.1	54.3 ± 1.2	55.93 ± 1.1
LT-integerated FL Methods										
Focal-FL (Lin et al., 2017)	95.3 ± 1.3	73.6 ± 0.1	2.7 ± 2.7	57.20 ± 1.3	58.63 ± 1.2	84.8 ± 1.3	71.3 ± 1.5	1.5 ± 2.0	52.54 ± 1.1	54.18 ± 1.0
LDAM-FL (Cao et al., 2019)	95.4 ± 0.5	72.2 ± 1.1	5.7 ± 3.9	57.77 ± 1.4	59.03 ± 1.3	86.9 ± 2.8	70.9 ± 1.2	4.7 ± 4.6	54.16 ± 1.4	55.61 ± 1.4
BSM-FL (Ren et al., 2020)	93.2 ± 1.5	74.6 ± 2.6	9.1 ± 3.7	58.92 ± 0.6	60.28 ± 0.7	89.6 ± 3.9	68.7 ± 3.0	16.4 ± 5.4	58.24 ± 1.2	59.15 ± 1.3
Label Distribution Skew FL										
FedLC (Zhang et al., 2022)	96.5 ± 0.4	75.3 ± 2.5	7.4 ± 5.5	59.73 ± 1.8	61.08 ± 1.7	95.8 ± 0.6	73.1 ± 2.4	6.6 ± 4.1	58.51 ± 1.5	59.78 ± 1.5
Federated Long-Tailed Methods (Fed-LT)										
CReFF (Shang et al., 2022)	95.1 ± 0.8	72.0 ± 1.5	2.6 ± 1.8	56.53 ± 1.4	57.88 ± 1.4	89.3 ± 0.7	70.1 ± 1.6	9.0 ± 4.5	56.12 ± 1.3	57.34 ± 1.2
FedFree (ours)	94.3 ± 1.2	72.9 ± 1.0	15.9 ± 2.7	**61.05 ± 0.3**	**62.08 ± 0.2**	93.0 ± 0.9	72.5 ± 2.6	16.2 ± 1.3	**60.57± 1.1**	**61.61 ± 1.0**

can offer a compelling starting point, increasing performance by 4.4% on "Acc" of ISIC-LT 1:500 with the baseline, BSM (Ren et al., 2020). Moreover, initialization with pre-trained weights can further increase our FoPro-KD ("Ours") performance by 5.0% on "Acc" on the ISIC-LT 1:500. Our method conserves the discriminative distribution capability while training through latent projections with EKD and input manipulation with FPG, which is vital for LT problems. On the HyperKvasir dataset, we observed that the performance of all methods improves when using the same optimization objective as in (Galdran et al., 2021), namely stochastic gradient descent (SGD) with a learning rate of 0.01 and a cosine annealing scheduler. We present additional experiments conducted on the HyperKvasir dataset in Figure 13. In (a), we observe that SGD with a learning rate of 0.1 and cosine annealing scheduler was optimal for random initialization, outperforming the setting proposed in (Galdran et al., 2021) and increasing the baseline performance (Ren et al., 2020) by 2.5% in terms of "B-Acc". However, when using ImageNet initialization with SGD 0.1, we observe in (b) a negative impact on performance, possibly due to the high learning rate distorting the pre-trained features. Nevertheless, our proposed Effective Knowledge Distillation (EKD) approach demonstrates its benefits on both random and pre-trained initialization in (b), with a minimal performance drop of 0.61% on "Ours" compared to a 1.65% drop on the best-performing baseline (Ren et al., 2020) in terms of "B-Acc". Furthermore, we can notice in (b) that "Ours" and the BKD approach (Zhang et al., 2023) are minimally affected by random and pre-trained model initialization. This can be attributed to the utilization of knowledge distillation, which helps maintain stability during training and prevents deviation in the target model's performance. Additionally, our results showed that SGD with a learning rate of 0.1 converged to an optimal point more effectively than most methods when using Adam. However, when initializing with pre-trained weights and using Adam, we achieved a mean of "B-acc" of 69.9%. Overall, our proposed method consistently outperformed other approaches across different weight initialization and optimization strategies, demonstrating its effectiveness.

*5.2. **FedFree***

In this section, we present the results of our proposed method, FedFree, in the federated learning setting. First, we start by comparing FedFree with various federated learning methods on the gastrointestinal dataset. Then, we shorten the benchmark to compare it on the federated skin-attribute setting.

5.2.1. Performance on HyperKvasir-FL

FL-Methods (Li et al., 2021, 2020; McMahan et al., 2017). One simple solution for federated long-tailed learning is to directly apply existing FL methods to our setting. To this end, we compare our methods with state-of-the-art FL methods, including FedAvg (McMahan et al., 2017), FedProx (Li et al., 2020), and MOON (Li et al., 2021), under the same setting. As shown in Table 8, we find that our method outperforms the best existing FL method MOON by 2.85% and 5.68% on "B-Acc" in both Type 1 and Type 2 settings, respectively. Notably, our FedFree achieves similar results with MOON (Li et al., 2021) on the head class while reaching large improvements on the tail class (11.9% on Type 1 and 10.71% on Type 2), showing that our FedFree can tackle LT distribution under FL more effectively. The limited results could be attributed to the use of local empirical risk minimization in MOON (Li et al., 2021). However, even when we applied a balanced risk minimization (BRM) in MOON, our method still outperformed it (60.69% vs. 62.08% on B-Acc for Type 1); we provide ablation results in Table 11.

LT integrated FL methods (Cao et al., 2019; Lin et al., 2017; Ren et al., 2020). To design FL methods for local clients with long-tailed distribution, a straightforward idea is to directly use LT methods in each local client and then use an FL framework such as FedAvg to obtain the final results. From Table 8, we can notice the LT methods utilizing an FL framework have produced

limited results in tail class classification, primarily due to the extreme client drifting phenomenon. Please note that in the FL, FedFree does not focus on designing specific long-tailed training for each local client. Instead, the DLMA module enables the global server to effectively aggregate the model parameters from long-tailed distributed local clients. As a result, our FedFree can successfully capture the tail class with a 6.84% tail accuracy gain on Type-1 with lower variance than the best-performing method BSM-FL (Ren et al., 2020). Notably, our method consistently outperforms the best-performing LT method on the "B-Acc" with a lower variance (improvement of 1.8% on Type-1 and 2.46% on Type-2).

Label-Skew FL FedLC (Zhang et al., 2022), inspired by (Cao et al., 2019), proposes a loss function to address label distribution skewness by locally calibrating logits and reducing local bias in learning. Their modification yields compelling performance. Nevertheless, our method surpasses them in both Type 1 and Type 2, achieving improvements of 1.0% and 1.83% in terms of balanced accuracy ("B-Acc"), respectively. Remarkably, our method effectively captures the tail classes with reduced variance in both Type 1 and Type 2, exhibiting improvements of 8.5% and 9.6%, respectively, while experiencing only a minor drop in performance for the head classes (96.5% vs 94.3% for Type 1 and 95.8% vs 93.0% for Type 2).

Fed-LT methods (Shang et al., 2022) We compare our method with the state-of-the-art Fed-LT method CReFF (Shang et al., 2022). CReFF, as proposed by (Shang et al., 2022), involves a method of re-training the classifier by utilizing learnable features on the server at each communication round, holding an equal treatment of all clients' models. However, this technique fails to accommodate inter-client intra-class variations which could arise. From Table 8, we can notice that FedAvg with local LT such as BSM-FL (Ren et al., 2020) can outperform CReFF (Shang et al., 2022) on the HyperKvasir dataset in both Type-1 and Type-2 by 2.4% and 1.8% on "B-Acc", respectively. Our comparative analysis illustrates that FedFree consistently outperforms CReFF in both Type-1 and Type-2 by 4.2% and 4.27% on "B-Acc", respectively, by addressing client drifting issues with our proposed FLKD (Fast and Convenient Local Model) and DLMA (Robust Estimation for the Global Model) innovations.

5.2.2. Performance on ISIC-FL Attribute

We evaluate the best-performing and competitive methods in the Fed-LT experiments with the ISIC-FL attribute dataset to shorten the benchmark. While previous studies neglect weight initialization to provide better convergence analysis as pre-trained weights are architecture dependent. Recently, Nguyen et al. (2023) and Chen et al. (2023) studied the impact of pre-training initialization on reducing the data and system heterogeneity in FL. We present in Table 9 the results of the most competitive methods with and without weight initialization on the ISIC-FL attribute setting. FedLC (Zhang et al., 2022) demonstrates compelling performance to address label skewness in Hyperkvasir-FL. Nevertheless, it falls short in accommodating attribute heterogeneity in ISIC-FL due to its local learning focus. Our method consistently outperforms FedLC (Zhang et al., 2022) with a notable improvement of 9.4% and 6.1% in terms of the averaged balanced accuracies ("Avg") and balanced accuracy ("B-Acc") respectively when clients' model weights are not available. When the client's model weights are available and initialized with ImageNet weights, the improvements are 10.2% and 1.79% on the "Avg" and "B-acc" respectively. Furthermore, our method exhibits a performance gain of 6.3% and 4.1% on "Avg" and "B-Acc" compared to the baseline (Ren et al., 2020) respectively when clients' models are not available, and 2.5% and 1.8 % on "Avg" and "B-Acc" respectively when client's model weights are available.

Table 9: Experimental Results on ISIC-FL-Attribute Split

Method	Metrics					
	HAM-1000		ISIC-19		All	
	Light	Dark	Light	Dark	Avg	B-Acc
	w/o Weight Initialization					
FedLC (Zhang et al., 2022)	50.09	54.99	61.78	41.20	52.02	57.33
BSM-FL (Ren et al., 2020)	54.79	56.61	62.38	46.58	55.09	59.40
FedFree w/o DLMA	55.09	63.28	62.89	52.33	58.39	60.36
FedFree (ours)	60.62	65.69	65.43	53.72	**61.37**	**63.45**
	w/ ImageNet Weight Initialization					
FedLC (Zhang et al., 2022)	66.75	37.80	74.25	79.40	64.55	71.39
BSM-FL (Ren et al., 2020)	66.93	68.90	74.53	78.55	72.23	72.15
FedFree (ours)	69.04	74.97	75.74	79.26	**74.75**	**73.18**

Table 10: Ablation of EKD and DLMA on HyperKvasir Type-2.

	EKD	DLMA	Metrics	
			All (%)	B-Acc (%)
Baseline (Ren et al., 2020)	×	×	58.24 ± 1.2	59.15 ± 1.3
w/ EKD	✓	×	59.26 ± 1.2	60.19 ± 1.1
w/ EKD + DLMA	✓	✓	**60.57 ± 1.1**	**61.61 ± 1.0**

5.2.3. Ablation

Effectiveness of EKD and DLMA. As shown in Table 10, applying the EKD to the baseline can enhance the "All" and "B-Acc" via 1.02% and 1.04%. With both EKD and DLMA, the performance is further improved to the best via 2.33% and 2.46% on "All" and "B-Acc", respectively. DLMA utilizes bias-free frozen generalizable representations to incorporate the inter-client intra-class characteristics in FL and combine it with the distillation belief of EKD (how well the information of free lunch models has been distilled to local clients). This combination helps in capturing client-specific models in the aggregation step.

Local Learning matters in FL Similarity to prior FL studies (Chen and Chao, 2022; Mendieta et al., 2022), we show that local learning matters in FL. We apply

Table 11: FL methods with local BRM.

Method	All		B-Acc	
	Type-1	Type-2	Type-1	Type-2
FedAvg	58.92	58.24	60.28±0.6	59.15±1.3
FedProx	59.37	58.86	60.47±1.3	59.64±2.0
Moon	59.45	58.72	60.69±0.9	59.66±0.8
FedFree (ours)	**61.05**	**60.57**	**62.08±0.2**	**61.61±1.4**

our baseline, BSM (Ren et al., 2020), as a local balanced risk minimization (BRM) with different federated learning algorithms. We can notice BRM can boost the performance of different federated methods. However, the best performance is achieved by our **FedFree** with 1.39% and 1.95% improvements on the "B-Acc" on Type-1 and Type-2 than the best performance FL method, MOON (Li et al., 2021). Although we use our LT estimations and findings to boost the FL framework, our framework can be further boosted via local LT re-sampling techniques (Ju et al., 2021, 2022) using our proposed distillation belief or via classifier-retraining (Kang et al., 2020) as shown in Table 12.

Table 12: Using a plug-in cRT on methods on Type-2.

Method + cRT	All	B-Acc
Decoupling (Kang et al., 2020)	54.21	55.6
BSM-FL (Ren et al., 2020)	62.67	63.11
FedFree	**65.05**	**65.11**

6. Discussion

Rare disease classification is a crucial aspect of medical imaging, and leveraging publicly available pre-trained models can potentially improve the diagnosis and representations of these diseases. Existing work in this area often regularizes training on synthesizing worst-case scenarios and extracting the knowledge using closed-set datasets without fully exploiting the generalization capabilities of widely known pre-trained models. Although some studies have explored effective prompting techniques for these models, their approaches are often limited to high-level features and prompt engineering without a deep understanding of how these "free lunch" encoders work, or how their representations can be further enhanced through a fundamental understanding of DNNs. In this work, we address this gap by investigating an intuitive phenomenon that has been widely neglected in the community: explicitly asking the pre-trained model what it wants, conditional on a cross-task medical input data, in order to gain insights into the learning dynamics of these models for effective representation learning. Through our method, we successfully demonstrate and leverage this phenomenon, shedding light on the inner workings of these models' frequency patterns and their behavior toward representation learning. Additionally, while skin lesions and gastrointestinal images can be considered out-of-distribution data for the free lunch model, there are extreme cases in medical imaging, such as X-rays and MRIs, which may require further exploration. Future research should aim to bridge the gap between natural image and medical imaging domains to enhance our understanding of the billions of parameters utilized in pre-trained models released yearly. On the other hand, in decentralized training, FedFree offered compelling performance in measuring the local bias and correcting it with the same discriminative consistent model. Nevertheless, theoretical proofs rather than empirical evaluations are needed for better convergence analysis to address nonvanishing terms and unlock further improvements.

7. Conclusion

In conclusion, we propose two frameworks to address long-tailed medical image classification tasks: FoPro-KD for centralized training and FedFree for decentralized training. FoPro-KD effectively compresses knowledge from publicly available pre-trained models into smaller target models. Future research can focus on exploring the generalization capabilities of pre-trained models and developing compression methods for medical imaging tasks. On the other hand, FedFree introduces a Federated Long-tailed Learning framework for decentralized training. Fed-Free promotes consistent learning in a decentralized setting with a dynamic aggregation method to effectively integrate inter-client variations. Both frameworks utilize the pre-trained model's knowledge to smaller target models for medical tasks that can be particularly useful in clinical settings where affordable AI is needed. Overall, our two proposed frameworks and findings represent a promising direction for addressing long-tailed classification problems and transfer learning in medical imaging.

8. Acknowledgments

I would like to express my sincere gratitude to Dr. Robert for his support and guidance throughout my thesis. Also, I would like to thank Dr. Xiaomeng Li for her supervision and hosting. I would also like to extend my thanks to Dr. Huzahu Fu for providing valuable insights and perspectives on the federated learning setting. Also, I would like to thank Hualiang Wang for meaningful discussions in the long-tailed learning.

References

Bai, J., Yuan, L., Xia, S.T., Yan, S., Li, Z., Liu, W., 2022. Improving vision transformers by revisiting high-frequency components, in: European Conference on Computer Vision.

Bevan, P.J., Atapour-Abarghouei, A., 2022. Detecting melanoma fairly: Skin tone detection andÂ debiasing forÂ skin lesion classification, in: Kamnitsas, K., Koch, L., Islam, M., Xu, Z., Cardoso, J., Dou, Q., Rieke, N., Tsaftaris, S. (Eds.), Domain Adaptation and Representation Transfer, Springer Nature Switzerland, Cham. pp. 1–11.

Borgli, H., Thambawita, V.L., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., Johansen, D., Griwodz, C., Stensland, H.K., Garcia-Ceja, E., Schmidt, P.T., Hammer, H.L., Riegler, M., Halvorsen, P., de Lange, T., 2019. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific Data 7.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T., 2019. Learning imbalanced datasets with label-distribution-aware margin loss, in: Advances in Neural Information Processing Systems.

Chen, C., Li, Z., Ouyang, C., Sinclair, M., Bai, W., Rueckert, D., 2022a. MaxStyle: Adversarial style composition for robust medical image segmentation, in: MICCAI. arXiv:2206.01737.

Chen, D., Mei, J.P., Zhang, H., Wang, C., Feng, Y., Chen, C., 2022b. Knowledge distillation with the reused teacher classifier, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11933–11942.

Chen, D., Mei, J.P., Zhang, Y., Wang, C., Wang, Z., Feng, Y., Chen, C., 2021a. Cross-layer distillation with semantic calibration. Proceedings of the AAAI Conference on Artificial Intelligence 35, 7028–7036. URL: https://ojs.aaai.org/index.php/AAAI/article/view/16865, doi:10.1609/aaai.v35i8.16865.

Chen, G., Peng, P., Ma, L., Li, J., Du, L., Tian, Y., 2021b. Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 458–467.

Chen, H.Y., Chao, W.L., 2022. On bridging generic and personalized federated learning for image classification, in: ICLR. URL: https://openreview.net/forum?id=I1hQbx10Kxn.

Chen, H.Y., Tu, C.H., Li, Z., Shen, H.W., Chao, W.L., 2023. On the importance and applicability of pre-training for federated learning, in: The Eleventh International Conference on Learning Representations. URL: https://openreview.net/forum?id=fWWFv--P0xP.

Chen, Z., Liu, S., Wang, H., Yang, H.H., Quek, T.Q.S., Liu, Z., 2022c. Towards federated long-tailed learning. ArXiv abs/2206.14988.

Combalia, M., Codella, N.C.F., Rotemberg, V.M., Helba, B., Vilaplana, V., Reiter, O., Halpern, A.C., Puig, S., Malvehy, J., 2019. Bcn20000: Dermoscopic lesions in the wild. ArXiv abs/1908.02288.

Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S., 2019. Class-balanced loss based on effective number of samples, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9268–9277.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255.

Ding, X., Liu, Z., Li, X., 2022. Free lunch for surgical video understanding by distilling self-supervisions, in: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (Eds.), MICCAI 2022, Springer. pp. 365–375.

Dong, B., Zhou, P., Yan, S., Zuo, W., 2023. LPT: Long-tailed prompt tuning for image classification, in: ICLR. URL: https://openreview.net/forum?id=8pOVAeo8ie.

Fang, G., Song, J., Wang, X., Shen, C., Wang, X., Song, M., 2021. Contrastive model inversion for data-free knowledge distillation. arXiv preprint arXiv:2105.08584 .

Galdran, A., Carneiro, G., González Ballester, M.A., 2021. Balanced-mixup for highly imbalanced medical image classification, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, Cham. pp. 323–333.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y., 2014. Generative adversarial nets, in: NIPS.

Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M., 2020. Bootstrap your own latent a new approach to self-supervised learning, in: NeurIPS, Curran Associates Inc., Red Hook, NY, USA.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.B., 2020. Momentum contrast for unsupervised visual representation learning. CVPR , 9726–9735.

Hu, S., Liao, Z., Xia, Y., 2022. Prosfda: Prompt learning based source-free domain adaptation for medical image segmentation. arXiv preprint arXiv:2211.11514 .

Huang, J., Guan, D., Xiao, A., Lu, S., 2021. Rda: Robust domain adaptation via fourier adversarial attacking. arXiv preprint arXiv:2106.02874 .

Jee Cho, Y., Wang, J., Joshi, G., 2022. Towards understanding biased client selection in federated learning, in: Camps-Valls, G., Ruiz, F.J.R., Valera, I. (Eds.), Proceedings of The 25th International Conference on Artificial Intelligence and Statistics, PMLR. pp. 10351–10375.

Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N., 2022. Visual prompt tuning, in: European Conference on Computer Vision (ECCV).

Ju, L., Wang, X., Wang, L., Liu, T., Zhao, X., Drummond, T., Mahapatra, D., Ge, Z., 2021. Relational subsets knowledge distillation for long-tailed retinal diseases recognition, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, Cham. pp. 3–12.

Ju, L., Wu, Y., Wang, L., Yu, Z., Zhao, X., Wang, X., Bonnington, P., Ge, Z., 2022. Flexible sampling forÂ long-tailed skin lesion classification, in: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, Springer Nature Switzerland, Cham. pp. 462–471.

Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y., 2020. Decoupling representation and classifier for long-tailed recognition, in: ICLR.

Kim, M., Li, D., Hospedales, T., 2023. Domain generalisation via domain adaptation: An adversarial fourier amplitude approach, in: ICLR. URL: https://openreview.net/forum?id=7IGOwsTND7w.

Kobayashi, T., 2021. Group softmax loss with discriminative feature grouping. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV) , 2614–2623.

Kumar, A., Raghunathan, A., Jones, R.M., Ma, T., Liang, P., 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution, in: ICLR. URL: https://openreview.net/forum?id=UYneFzXSJWh.

Li, Q., He, B., Song, D., 2021. Model-contrastive federated learning, in: CVPR.

Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V., 2020. Federated optimization in heterogeneous networks, in: Dhillon, I., Papailiopoulos, D., Sze, V. (Eds.), Proceedings of Machine Learning and Systems, pp. 429–450.

Li, Z., Shang, X., He, R., Lin, T., Wu, C., 2023. No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier. ArXiv abs/2303.10058.

Lin, T.Y., Goyal, P., Girshick, R.B., He, K., Dollár, P., 2017. Focal loss for dense object detection. ICCV , 2999–3007.

Liu, Q., Yang, H., Dou, Q., Heng, P.A., 2021. Federated semi-supervised medical image classification via inter-client relation matching, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, Cham. pp. 325–335.

Loshchilov, I., Hutter, F., 2017. SGDR: Stochastic gradient descent with warm restarts, in: ICLR.

Makino, T., Jastrzebski, S., Oleszkiewicz, W., Chacko, C., Ehrenpreis, R., Samreen, N., Chhor, C., Kim, E., Lee, J., Pysarenko, K., Reig, B., Toth, H., Awal, D., Du, L., Kim, A., Park, J., Sodickson, D.K., Heacock, L., Moy, L., Cho, K., Geras, K.J., 2020. Differences between human and machine perception in medical diagnosis. Scientific Reports 12.

McMahan, H.B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A., 2017. Communication-efficient learning of deep networks from decentralized data, in: AISTATS.

Mendieta, M., Yang, T., Wang, P., et al., 2022. Local learning matters: Rethinking data heterogeneity in federated learning, in: CVPR, pp. 8397–8406.

Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. URL: http://arxiv.org/abs/1411.1784. cite arxiv:1411.1784.

Mu, X., Shen, Y., Cheng, K., Geng, X., Fu, J., Zhang, T., Zhang, Z., 2021. Fedproc: Prototypical contrastive federated learning on non-iid data. Future Gener. Comput. Syst. 143, 93–104.

Nguyen, J., Wang, J., Malik, K., Sanjabi, M., Rabbat, M., 2023. Where to begin? on the impact of pre-training and initialization in federated learning, in: The Eleventh International Conference on Learning Representations. URL: https://openreview.net/forum?id=Mpa3tRJFBb.

Oh, J., Kim, S., Yun, S.Y., 2022. FedBABU: Toward enhanced representation for federated image classification, in: International Conference on Learning Representations. URL: https://openreview.net/forum?id=HuaYQfggn5u.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H.Q., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M.G., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P., 2023. Dinov2: Learning robust visual features without supervision. ArXiv abs/2304.07193.

Pereyra, G., Tucker, G., Chorowski, J., Kaiser, L., Hinton, G., 2017. Regularizing neural networks by penalizing confident output distributions. URL: https://openreview.net/forum?id=HkCjNI5ex.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning.

Reinke, A., Christodoulou, E., Glocker, B., et al., 2022. Metrics reloaded - a new recommendation framework for biomedical image analysis validation, in: Medical Imaging with Deep Learning.

Ren, J., Yu, C., Sheng, S., Ma, X., Zhao, H., Yi, S., Li, H., 2020. Balanced meta-softmax for long-tailed visual recognition, in: Proceedings of Neural Information Processing Systems(NeurIPS).

Shang, X., Lu, Y., Huang, G., Wang, H., 2022. Federated learning on heterogeneous and long-tailed data via classifier re-training with federated features, in: Raedt, L.D. (Ed.), IJCAI, pp. 2218–2224.

Shuai, X., Shen, Y., Jiang, S., Zhao, Z., Yan, Z., Xing, G., 2022. Balancefl: Addressing class imbalance in long-tail federated learning, in: 2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), pp. 271–284. doi:10.1109/IPSN54338.2022.00029.

Tang, K., Tao, M., Qi, J., Liu, Z., Zhang, H., 2022. Invariant feature learning for generalized long-tailed classification, in: ECCV, p. 709–726.

Tschandl, P., Rosendahl, C., Kittler, H., 2018. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data 5.

Wang, Y., Cheng, J., Chen, Y., Shao, S., Zhu, L., Wu, Z., Liu, T., Zhu, H., 2023. Fvp: Fourier visual prompting for source-free unsupervised domain adaptation of medical image segmentation. ArXiv abs/2304.13672.

Wicaksana, J., Yan, Z., Cheng, K.T., 2023. Fca: Taming long-tailed federated medical image classification by classifier anchoring. ArXiv abs/2305.00738.

Wu, N., Yu, L., Yang, X., Cheng, K.T., Yan, Z., 2022. Federated learning with imbalanced and agglomerated data distribution for medical image classification.

Yang, C., Guo, X., Chen, Z., Yuan, Y., 2022. Source free domain adaptation for medical image segmentation with fourier style mining. Medical Image Analysis 79, 102457. URL: https://www.sciencedirect.com/science/article/pii/S1361841522001049, doi:https://doi.org/10.1016/j.media.2022.102457.

Yang, W., Chen, D., Zhao, H., Meng, F., Zhou, J., Sun, X., 2023. Integrating local real data with global gradient prototypes for classifier re-balancing in federated long-tailed learning. ArXiv abs/2301.10394.

Ye, J., Ji, Y., Wang, X., Gao, X., Song, M., 2020. Data-free knowledge amalgamation via group-stack dual-gan, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12513–12522. doi:10.1109/CVPR42600.2020.01253.

Yu, A., Yang, Y., Townsend, A., 2023. Tuning frequency bias in neural network training with nonuniform data, in: ICLR. URL: https://openreview.net/forum?id=oLIZ2jGTiv.

Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2018. mixup: Beyond empirical risk minimization, in: ICLR. URL: https://openreview.net/forum?id=r1Ddp1-Rb.

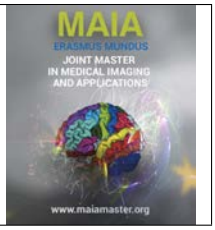
Zhang, J., Li, Z., Li, B., Xu, J., Wu, S., Ding, S., Wu, C., 2022. Federated learning with label distribution skew via logits calibration, in: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (Eds.), Proceedings of the 39th International Conference on Machine Learning, PMLR. pp. 26311–26329. URL: https://proceedings.mlr.press/v162/zhang22p.html.

Zhang, S., Chen, C., Hu, X., Peng, S., 2023. Balanced knowledge distillation for long-tailed learning. Neurocomputing 527, 36–46. URL: https://www.sciencedirect.com/science/article/pii/S0925231223000711, doi:https://doi.org/10.1016/j.neucom.2023.01.063.

Zhao, G., Yang, W., Ren, X., Li, L., Sun, X., 2021. Well-classified examples are underestimated in classification with deep neural networks, in: AAAI Conference on Artificial Intelligence.

Medical Imaging and Applications

Master Thesis, June 2023



Domain generalization for multiple sclerosis lesion segmentation in brain MRI

Rachika Elhassna Hamadache, Arnau Oliver, Xavier Lladó

Research institute of Computer Vision and Robotics (ViCOROB), Universitat de Girona, Catalonia, Spain

Abstract

Multiple sclerosis (MS) is a progressive disease of the central nervous system, characterized by lesions of different shapes and sizes. Recently, computer-aided diagnosis (CAD) systems based on deep learning attained remarkable results in segmenting MS lesions in magnetic resonance imaging (MRI). However, when tested on images from different domains (ie, scans acquired from different MRI scanners and protocols), these systems show an important drop in performance, hence become unreliable. In this master thesis, we explore several approaches based on state-of-the-art strategies in terms of segmentation and domain generalization (DG) to tackle this domain shift problem. Starting from a 3D residual UNet (ResUNet) architecture, we incorporate some recently proposed modules for enhancing feature representation learning. Moreover, variational autoencoder based architectures are also considered to evaluate the impact of their feature regularization ability on DG, as well as transformers-based networks to assess their robustness and efficiency in medical imaging segmentation. All models are trained on the recent Shifts challenge 2023 dataset, and the best model is further tuned using more images from an in-house dataset of the Vall d'Hebron University Hospital (Barcelona). The obtained results show that the ResUNet trained on relevant feature representations achieves the best performances among the studied methods, but remains weak in detecting all lesions in unseen domains, due to the limited training data and the difficulty of the MS lesion segmentation task itself.

Keywords: Multiple sclerosis, MRI, lesion segmentation, deep learning, domain generalization, Shifts challenge.

1. Introduction

Multiple sclerosis (MS) is a progressive and incurable inflammatory-demyelinating disease of the central nervous system (CNS) that negatively alters individuals' lives (Malinin et al., 2022). It has reached 2.8 million cases in 2020 (Walton et al., 2020) and became the most common non-traumatic neurological disease diagnosed among young adults (Lladó et al., 2012). The pathologic hallmark of MS includes multiple focal areas of myelin loss and inflammation (referred to as plaques or lesions), axonal loss and gliosis scattered within the CNS (Lladó et al., 2012; Popescu et al., 2013).

Currently, magnetic resonance imaging (MRI) is a fundamental technique used to characterize and quantify MS lesions, which is essential for the disease diagnosis, progression tracking and evaluation of the treatment's efficacy (Valverde et al., 2019). MRI protocols mainly include fluid-attenuated inversion recovery (FLAIR), T2-w and T1-w modalities for their complementary contrasts and high sensitivity in detecting the presence of lesions (Ackaouy et al., 2020).

Even though visual lesion inspection is practically feasible, the manual delineation of MS lesions is highly challenging and time-consuming for the large number of MRI slices to assess per patient, as well as prone to intra- and inter- expert variability (Salem et al., 2022). This has led to an increasing interest in the development of automated MS lesion segmentation methods, making it an active field of research.

In fact, recent years have shown the emergence of various image processing methods based on deep learning (DL). In particular, convolutional neural networks (CNN) architectures have demonstrated remarkable performances in different fields, including medical imaging, which helped in building better computer-aided diagnosis (CAD) systems (Kamraoui et al., 2022; Valverde et al., 2019).

Nonetheless, in machine learning (ML) and DL strategies, it is commonly assumed that the training and

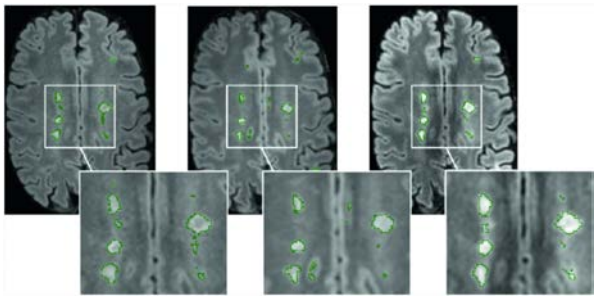


Figure 1: MRI scans from one patient in three 3T scanners with automated lesion segmentations in green. From left to right: Philips Achieva, Siemens Skyra and GE Discovery (Billast et al., 2020).

testing data are independent and identically distributed (Malinin et al., 2022), meaning that they come from the same data distribution. In real life medical applications, this assumption does not always hold, especially for MRI which can exhibit high or subtle variations across different sites and medical centers due to the acquisition protocols, MRI scanners, softwares and individuals (Ackaouy et al., 2020). Figure 1 shows an example of these variations found when the same patient underwent a brain scan from different MRI scanners.

These biases result in a poor generalization of the trained models when applied to new unseen target images, hence a decrease in their performances and applicability (as seen in the segmentations of Figure 1). This is known as the domain shift problem (Kamraoui et al., 2022).

For the MS lesion segmentation task, the limited availability of public datasets that describe the heterogeneity of the pathology, along with the various MRI domains, make this task even more challenging and require more robust and reliable strategies to tackle it. Several international challenges have been organized for this matter, such as ISBI (Carass et al., 2017), MSSEG (Commowick et al., 2018) and the most recent one Shifts challenge 2023 (Malinin et al., 2022), which particularly aims to handle the domain shift problem. These challenges serve as platforms to benchmark and foster collaboration among researchers to ultimately improve the reliability of the MS lesion segmentation techniques.

In this study, we start from a basic segmentation pipeline and explore DL based methods proposed in the recent literature on domain generalization (DG) to develop and discuss reliable approaches for segmenting MS lesions and potentially help in the early diagnosis and follow-up of the disease.

Similarly to Jin et al. (2021) but in a medical imaging context, we introduce a style normalization and restitution (SNR) module to a residual 3D UNet architecture to benefit from its style-normalized and task-relevant features. Other mechanisms such as attention and histogram matching are introduced to assess their impact on handling domain shifts. In addition, variational autoencoder (VAE) based architectures are also evaluated in terms of regularizing features to address DG, as well as transformers-based architectures to assess their robustness and efficiency in this segmentation task. The different proposed pipelines are trained on a relatively small annotated dataset (available from the Shifts challenge 2023 (Malinin et al., 2022)) and evaluated over 24 cases, allowing to analyse the impact of the different contributions introduced in this work.

The best performing pipelines were submitted to the online testing platform (Grand-Challenge) to be evaluated on a testing set of 74 out-of-domain (OOD) cases and compared with the state-of-the-art of the Shifts challenge 2023.

2. Literature review

Regarding the importance of enhancing the generalization ability of ML and DL methods, several related research topics have emerged, such as domain adaptation (DA) and DG, while others have been explored for this matter, like transfer learning and meta-learning (Wang et al., 2022).

DA occupies most of the literature for adapting specific target datasets on the source domain according to the applications. But in recent years, DG has received greater emphasis, as it aims to learn a model from one or several different but related domains that will generalize well on unseen target domains (Wang et al., 2022). These target data are totally unknown, not even unsupervised as in DA, which makes it more interesting, especially in the medical field where data is scarce and cannot include all possible variations for a better generalization.

According to Wang et al. (2022), the existing work on DG for computer vision can be categorized in the literature into three main groups:

1) Data manipulation
It is among the simplest and less computationally demanding way to increase both the quantity and diversity of the training data. It includes data augmentations and data generation through generative models, such as VAEs, generative adversarial networks (GAN) and the MixUp strategy. Examples of these approaches are the work of Somavarapu et al. (2020), which employs adaptive instance normalization (AdaIN) to achieve fast stylization to arbitrary styles, and Li et al. (2021a), which generates domains instead of samples via adversarial training, but remains highly complex and quite computationally expensive.
A much more simple approach is MixUp (Zhang et al., 2017), which consists of generating new samples or new features by performing linear interpolation between any two instances and between their labels.

2) Representation learning
It is the most popular approach in DG and can be divided into two main techniques:

- *Domain invariant representation-based DG*, whose goal is to reduce the representation discrepancy between multiple source domains to make the feature space more domain invariant and thus, make the learnt model have a better generalizability to unseen domains. In this division, we find domain adversarial neural networks (DANN) (Ganin et al., 2016), in which the discriminator is trained to distinguish the domains while the generator is trained to fool the discriminator to learn domain invariant feature representations. This idea was further adapted in a DG way by Li et al. (2018b).
  Other strategies focus on the normalization of features, such as in IBNNet (Pan et al., 2018) where instance normalization (IN), a task agnostic layer, is used alongside batch normalization (BN) to preserve discriminative information. Nam and Kim (2018) takes it one step further by replacing BN layers by batch-instance normalization (BIN) layers. Another interesting idea is found in Jin et al. (2021), where a SNR module was introduced. It consists of performing style normalization via IN, then by training a weighting vector, find the task-relevant features and add them back with the style-normalized features to restore helpful information and obtain better discrimination.

- *Feature disentanglement-based DG*, whose approaches consist of decomposing a feature representation into one that is domain invariant, and the other that is domain specific (Wang et al., 2022). An example of it is domain-invariant variational autoencoder (DIVA) (Ilse et al., 2020), that disentangles the features into domain information, class information, and other residual information. Another example is SagNets (Nam et al., 2021) in which they disentangle the style encodings from the class categories to better highlight the content.

3) Learning strategy
It focuses on exploiting the general learning strategy to promote the generalizability of the trained models. Among the works found in this category, Zhou et al. (2021) proposes Domain Adaptive Ensemble Learning (DAEL), which is composed of a shared CNN feature extractor across domains and multiple domain-specific classifier heads (experts). The DAEL aims to learn in a collaborative way, such that the experts teach the non-expert classifiers, and with that, encourage the ensemble to learn how to handle data from unseen domains. Li et al. (2018a) proposes meta-learning for domain generalization (MLDG) that splits the source domains data into meta-train and meta-test to simulate the shifts in domain and learn general representations. Other works focus on using gradient information to force the model to learn generalized representations, such as Shi et al. (2021), which aims to maximize the gradient inner product to align the gradients' directions that are assumed to be the same across domains.

Focusing back on the medical-related DG works, regardless of their categories, Li et al. (2021b) made use of histogram matching (HM) with an encoder-decoder (ED) architecture to achieve automatic left atrial segmentation from multi-center late gadolinium enhanced MRI. The paper shows that the simple HM managed to outperform other DG strategies that consisted of mutual information-based feature disentanglement (MIDNet) and pseudo-novel domain augmentation via random style transfer (RST-Net).

Hu et al. (2022) proposes a multi-source domain generalization model (DCAC) based on an ED architecture and domain and content adaptive convolutions for medical image segmentation. The general idea is to feed the globally average-pooled and concatenated feature maps of the encoder layers to a domain predictor that generates a domain code. This code is later used by a domain-aware controller to predict the parameters of a domain-adaptive head, and another content-aware controller predicts the parameters of a content-adaptive head, which is used to obtain the segmentation results.

As for MS lesion segmentation, Kamraoui et al. (2022) is among the very few papers that tackle this problem in a DG way. It presents DeepLesionBrain (DLB), a segmentation framework based on a spatially distributed strategy that uses a large group of overlapped compact 3D UNets, each one specialized in a certain region of the brain. This allows to produce robust predictions compared to an individual network. DLB also uses some data augmentations to increase training data variability, and a hierarchical specialization learning (HSL) strategy by pre-training a generic network over the whole brain, then use its weights as initialization to the locally specialized networks. This results in the network learning both generic and specific features extracted at global and local image levels respectively.

By reviewing the current state-of-the-art in DG applied to the medical field, very few works seem feasible for this project, regarding the limited available data for training and the difficulty of the task itself. Nonetheless, some ideas will be worth exploring. In this research, we start from a basic segmentation baseline and try to improve it by investigating the impact of some of the previously mentioned ideas in terms of robustness and generalizability.

3. Material and methods

3.1. Datasets

In this work, two datasets are available to explore DG in MS lesion segmentation: one from the Shifts chal-

Table 1: Meta information and splits of the Shifts challenge 2023 and VH datasets. Scanners are: Siemens Verio, GE Discovery, Siemens Aera, Philips Ingenia, Philips Medical, Siemens Magnetom Trio and Siemens Tim Trio.

		Location	Scanner	Field	Resolution (mm^3)	Raters	*Train*	Dev_{in}	$Eval_{in}$	Dev_{out}	$Eval_{out}$
Shifts challenge 2023	**MSSEG-1**	Rennes	S Verio	3.0 T	0.50 x 0.50 x 1.10	7	8	2	5	0	0
		Bordeaux	GE Disc	3.0 T	0.47 x 0.47 x 0.90	7	5	1	2	0	0
		Lyon	S Aera	1.5 T	1.03 x 1.03 x 1.25	7	10	2	17	0	0
			P Ingenia	3.0 T	0.74 x 0.74 x 0.70						
	ISBI	Best	P Medical	3.0 T	0.82 x 0.82 x 2.20	2	10	2	9	0	0
	PubMRI	Ljubljana	S Mag	3.0 T	0.47 x 0.47 x 0.80	3	0	0	0	24	0
	Private	Lausanne	S Mag	3.0 T	1.00 x 1.00 x 1.20	2	0	0	0	0	74
VH	**Private**	Barcelona	S Tim	3.0 T	0.49 x 0.49 x 3.00	1	30	0	27	0	0

lenge 2023, the MS lesion segmentation track that took place from september 2022 to may 2023 (Malinin et al., 2022), and an in-house one from the Vall d'Hebron University Hospital of Barcelona (Valverde et al., 2019). Details on each of them are presented in the following subsections.

3.1.1. Shifts challenge 2023 dataset

The dataset for the MS lesion segmentation track is a combination of several publicly available datasets[1] (ISBI, MSSEG-1 and PubMRI) and one private dataset from the university of Lausanne, which is not released and kept for submission evaluation on the GrandChallenge platform.

The data consists of 3D FLAIR and T1-w brain scans already pre-processed using denoising (non-local mean denoising filter), skull stripping, bias field correction (N4ITK), registration to FLAIR and interpolation to 1mm isovoxel space (Malinin et al., 2022). As for the ground-truth (GT) masks, they were obtained from a consensus of multiple expert annotators, except for Best and Lausanne (single mask) (Malinin et al., 2022).

For the Shifts challenge 2023, the data was structured in a 'canonical partitioning' in order to have in-domain training, validation and testing (*train*, dev_{in} and $eval_{in}$ respectively), as well as OOD testing (dev_{out} and $eval_{out}$) which were also shifted relative to each other. The details on the characteristics and splits of the data can be found in Table 1. The partitioning was selected based on experiments, where ensembles of 5 UNet models were trained on each data location and tested on all the others to identify the one presenting the highest shift. Other experiments using trained models on all data except for one location at a time allowed to confirm that Ljubljana (PubMRI) presents the greatest shift, followed by the private dataset, hence were both chosen as the OOD data while the rest were considered as the in-domain (Malinin et al., 2022).

It is worth mentioning that the variations in the data are not exclusive to the locations or scanner types, but also related to the scanner strengths (1.5T and 3T), annotators' guidelines, resolution of the raw FLAIR scans and the size of lesions (OOD data presents more subjects with smaller lesions) (Malinin et al., 2022).

These conditions portrait real-world distributional shift which makes the data suitable for assessing the robustness and generalizability of MS lesion segmentation solutions.

3.1.2. Vall d'Hebron dataset

This local dataset is provided by the Vall d'Hebron (VH) University Hospital in Barcelona, Spain. The FLAIR (TR=9000 ms, TE=93 ms, TI=2500 ms, flip angle=120°, voxel size=$0.49\times0.49\times3mm^3$) and T1-w (TR=2300 ms, TE=2 ms, TI=900 ms, flip angle=9°, voxel size=$1\times1\times1.2mm^3$) images were acquired from a 3T Siemens scanner with a 12-channel phased-array head coil. The data was pre-processed following the work of Valverde et al. (2019), which included skull stripping, N3 bias field correction, co-registration to T1-w (FSL-FLIRT) and interpolation to 1mm isovoxel space. The dataset was randomly partitioned into train and test sets (details are shown in Table 1) and presents cases with relatively low lesion loads.

[1] Data were generated by participating neurologists in the framework of Observatoire Français de la Sclérose en Plaques (OFSEP), the French MS registry (Vukusic et al. 2020). They collect clinical data prospectively in the European Database for Multiple Sclerosis (EDMUS) software (Confavreux et al. 1992). MRI of patients were provided as part of a care protocol. Nominative data are deleted from MRI before transfer and storage on the Shanoir platform (Sharing NeurOImagingResources, shanoir.org).

Vukusic S, Casey R, Rollot F, Brochet B, Pelletier J, Laplaud D-A, et al. Observatoire Français de la Sclérose en Plaques (OFSEP): A unique multimodal nationwide MS registry in France. Mult Scler. 2020;26(1):118–22.

Confavreux C, Compston DAS, Hommes OR, McDonald WI, Thompson AJ. EDMUS, a European database for multiple sclerosis. J Neurol Neurosurg Psychiatry 1992; 55: 671-676.

Andrey Malinin, Andreas Athanasopoulos, Muhamed Barakovic, Meritxell Bach Cuadra, Mark JF Gales, Cristina Granziera, Mara Graziani, Nikolay Kartashev, Konstantinos Kyriakopoulos, Po-Jui Lu, Nataliia Molchanova, Antonis Nikitakis, Vatsal Raina, Francesco La Rosa, Eli Sivena, Vasileios Tsarsitalidis, Efi Tsompopoulou, Elena Volf. Shifts 2.0: Extending The Dataset of Real Distributional Shifts, arxiv preprint https://arxiv.org/abs/2206.15407

3.2. Network architectures

This subsection presents the main network architectures used in the evaluated approaches. The main objective is to design a robust MS lesion segmentation model that can handle domain shifts in terms of segmentation and detection sensitivity (lesion-wise). In addition, the model should ideally be able to comply with the constraints of the Shifts challenge 2023, which impose predictions within 800ms per input sample, and the respect of the data partitioning, such that models are build only using the data provided by the organizers (*train* and dev_{in} for training and validation purposes respectively) or using data that is publicly released.

For this task, 3D encoder-decoder based architectures with skip connections are preferred for their simplicity and performances. We will analyze their ability to extract relevant and regularized feature representations with the help of additional components described later in section (3.4). In what follows, we describe the main architectures used within this work.

3.2.1. UNet (baseline)

The organizing team of the Shifts challenge 2023 provided their baseline model as a starting point. It is based on a 3D UNet (Çiçek et al., 2016) architecture with 5 layers and a strided convolution at the beginning of each block (for down- and up- sampling), which immediately reduces the spatial dimension of the input data by 2 (as shown in Figure 2). The network is implemented in MONAI. Three of this model are trained with different seeds on 3D patches of 96x96x96 voxels, for a maximum of 300 epochs with early stopping and no dropout. The trained models' predictions on the OOD data (dev_{out}) are ensembled and the resulting performance will be taken as a reference for further improvements in the following approaches.

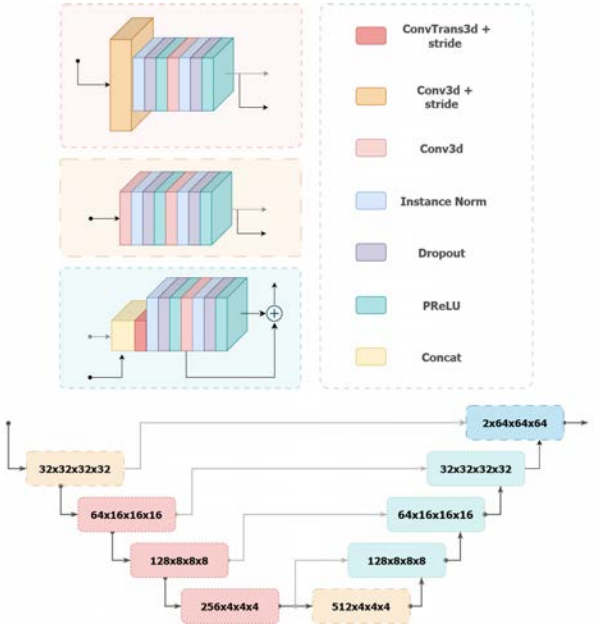


Figure 2: 3D UNet architecture as implemented in MONAI.

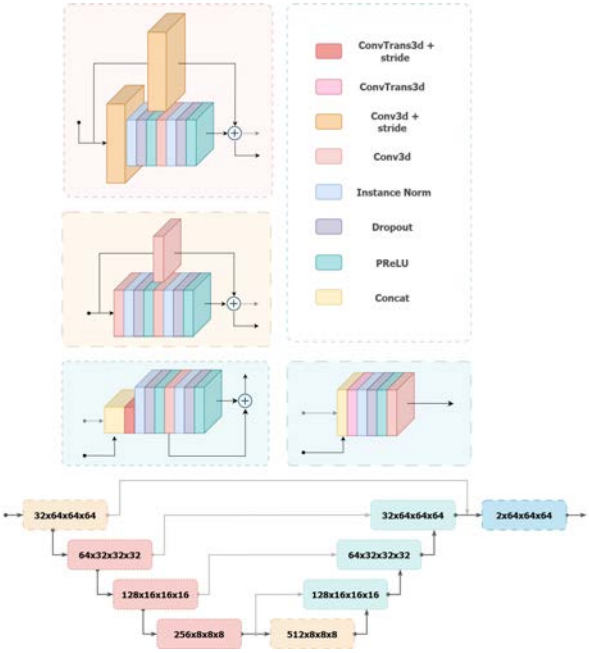


Figure 3: 3D ResUNet architecture as implemented in MONAI.

3.2.2. ResUNet

Residual UNet (ResUNet) (Kerfoot et al., 2019) is an enhanced version of the UNet architecture in which residual units are included to facilitate information flow and enhance training efficiency to better capture complex patterns and improve general performances and adaptability to unseen data.

The network consists of an ED of 5 layers with residual units, and connected through skip-connections. It uses convolutions and transpose convolutions with a stride of 2 for down-sampling and up-sampling the data respectively. This allows the network to learn optimal down-sampling and up-sampling operations while reducing the spatial dimension (Kerfoot et al., 2019). Parametric rectifying linear units (PReLU) are used to allow the network to learn the slop of the negative part of ReLU as a parameter, hence having a better activation and training. Dropout layers of 35% are added to avoid overfitting, and IN is applied to help generalization by removing instance-specific contrast information from the features. The overall architecture can be seen in Figure 3.

Having this model as an initial network, two additional modules are added to study their impact on DG: SNR (Jin et al., 2021) and attention gate (AG) (Oktay et al., 2018) modules. The details of each approach are explained in the subsection (3.4).

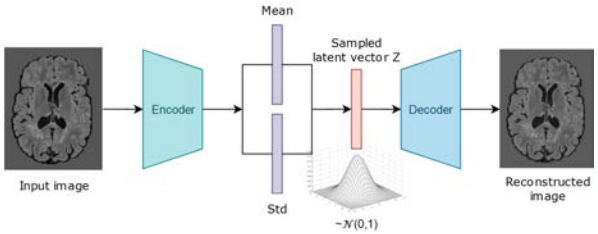


Figure 4: Graphical representation of a VAE.

3.2.3. VAE

VAEs (Kingma and Welling, 2013) are generative models based on an ED architecture and used for their efficient representation learning. The overall idea consists of first mapping the input data by the encoder to low-resolution and high-level feature representations, then this endpoint output is mapped to a latent space that corresponds to a low-dimentional space in which half represents the mean and the other half the standard deviation (std) of a distribution. Then, a sample is drawn from that latent space and used by the decoder to reconstruct the original input data (as seen in Figure 4). The training of such models is done by minimizing a VAE loss function composed of two terms: a reconstruction loss to make the ED reconstruction as performant as possible, and a regularization loss that makes the latent space distribution close to a standard normal distribution $\sim\mathcal{N}(0, 1)$. This is further detailed in the loss functions part of subsection (3.3.2).

With the hypothesis that VAEs learn to encode input data into a compact and meaningful latent representation, and following the successful work done by Myronenko (2019) and Li et al. (2020) in terms of 3D brain tumors segmentation from MRI, the SegResNetVAE (Myronenko, 2019) model (winner of BraTS2018 challenge) and a ResUNetVAE model are considered for this task to study the effect of VAEs on DG. The details of each approach are explained in the subsection (3.4).

3.2.4. Transformers

Following the success of transformers in NLP, Hatamizadeh et al. (2022b) proposed UNEt TRansformers (UNETR), a UNet-like architecture where the encoder is replaced by transformer blocks, with skip connections connecting them to the decoder, as seen in Figure 5. As in Vision Transformers (ViT), the images are separated into patches to be linearly projected into token embeddings and added with the positional embeddings.

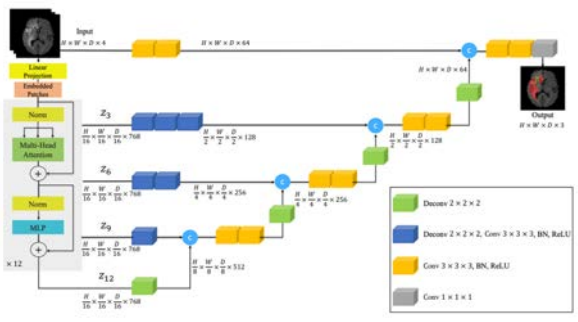


Figure 5: Overview of the UNETR architecture from Hatamizadeh et al. (2022b).

From the same authors, Hatamizadeh et al. (2022a) proposed SwinUNETR, which employs a Swin transformer block as the encoder, achieving by that impressive performances in some medical image segmentation works. Therefore, both networks are evaluated on their applicability in this MS lesion segmentation task.

3.3. Training details

This subsection presents some implementation and training details used in the studied strategies.

3.3.1. Data transforms

Following the challenge's baseline, the brain volumes are split into 3D patches of 64x64x64 voxels and normalized to zero mean and unit std. During training, 128 patches are randomly sampled from the original inputs such that they are centred on a lesion voxel 80% of the time. Other random transforms from MONAI are used: intensity shift and scale, flipping, rotation and affine transformation.

As for validation and inference, patches overlapping by 50% and 70% respectively are used with Gaussian weighting averaging to get the final segmentation predictions for the ResUNet based approaches. For the rest of the methods, the overlapping of patches was by 25% in validation.

For each of the studied approaches, 3 single models are trained (each one initialized with a different seed) to form a deep ensemble (by averaging their output probabilities) and increase their robustness. These probabilities are then thresholded to obtain the final per-voxel segmentation map. The threshold value remained unchanged from the challenge baseline (0.35) since experimenting with different values did not yield significant variations in the results.

3.3.2. Optimization

- Loss functions: the loss functions help guiding the learning process and optimization of a model by quantifying the difference between the true and predicted values. For this task, several loss functions have been used according to the model's needs:
  - Dice loss (Milletari et al., 2016), which measures the similarity between the predicted and ground truth segmentation masks by computing the overlap between the two.
  - Focal loss (Lin et al., 2017), a cross-entropy based loss that addresses class imbalance by focusing on challenging regions and minority classes during training.

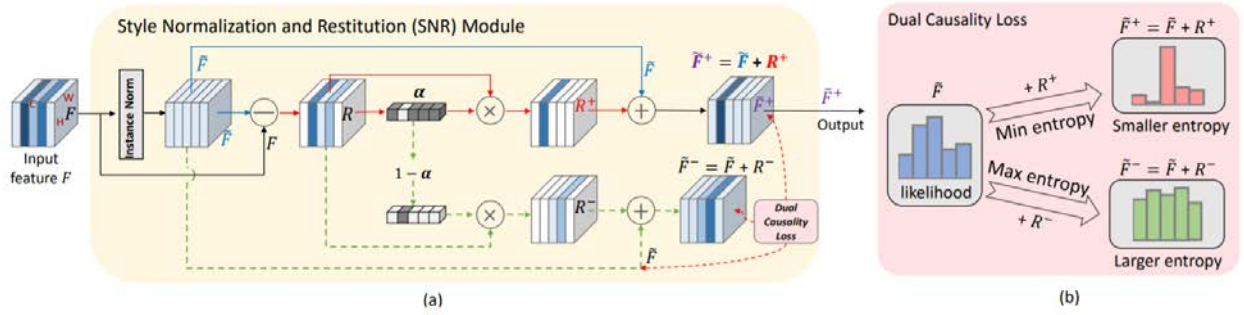


Figure 6: Overall structure of the SNR module from Jin et al. (2021). From left to right: SNR module and dual restitution loss constraint.

- VAE loss, which is the combination of two terms: the reconstruction loss and the regularization loss. The reconstruction loss between the original image and the reconstructed one encourages the VAE to generate outputs that are as close as possible to the original input (usually mean squared error: MSE). The regularization loss between the learned distribution of the latent space and the target distribution (usually the standard normal distribution) helps to regularize the latent space and ensures that it follows the chosen target distribution (Kullback-Leibler divergence loss: KLD).
- Dual restitution loss (Jin et al., 2021), as illustrated in Figure 6b, encourages the disentanglement of task-relevant and task-irrelevant features by comparing the discrimination capability of features before and after restitution. This will be further explained in the subsection (3.4).

- Optimizer: Adaptive moment estimation (Adam) (Kingma and Ba, 2014) optimizer is used with a ReduceOnPlateau learning rate (LR) scheduler to improve the model's training. The starting LR is $1e^{-3}$ and will be reduced by a factor of 2 if the validation loss does not improve after 7 epochs. This configuration is common for all the studied methods.

- Early stopping: all models are trained with 300 maximum epochs and an early stopping with a patience of 50 according to the validation loss (on dev_{in}). The patience was overly increased due to the relatively slow training of the models, thus to give more chances for the models to improve.

3.4. Evaluated approaches

This subsection presents the main experiments done in this project and trained on the FLAIR images of the Shifts challenge 2023 dataset (*train* and dev_{in}). They are devided into three main parts according to the network architecture they are based on.

3.4.1. ResUNet

3.4.1.1. ResUNet

As described in the previous subsection (3.2.2), the first approach consists of leveraging the available baseline by training a 5 layers 3D residual UNet from scratch on the available data. The model, as seen in Figure 3, is implemented in the MONAI library and the training was done with a batch size of 3 and a loss function that combines the dice loss and the focal loss.

3.4.1.2. ResUNet SNR

As introduced in the literature review (2), the SNR (Jin et al., 2021) module aims to enhance both the generalization and discrimination capabilities of any model by inserting the module after convolutional blocks. The detailed architecture of the module is shown in Figure 6a, and can be summarized as follows. First, IN is applied to the input features F to reduce their instance discrepancy and obtain some kind of style-normalized features $\tilde{F}$. Then, to counter the loss of discriminative information induced by IN, the next steps aim to restitute the task-relevant features from the residuals R by disentangling them into task-relevant R^+ and task-irrelevant R^- features through masking R by a learned channel attention vector a. With that, the task-relevant features R^+ are added back with the normalized features $\tilde{F}$ to restitute important information and obtain $\tilde{F}^+$.

For our segmentation task, the SNR module is added after each convolutional block of the encoder, to obtain better and robust features before feeding them to the decoder. For training this model, the batch size was 4 and the loss function combined the dice loss, the focal loss and the dual restitution loss (Jin et al., 2021). This latter is illustrated in Figure 6b and follows the intuition that after restitution, the enhanced features $\tilde{F}^+$ become more discriminative and thus, when the feature vector of each spatial position of $\tilde{F}^+$ is passed to a fully connected layer of K nodes followed by a softmax (K being the number of classes, here 2), the predicted class likelihood should have a smaller entropy. On the other hand, when adding the task-irrelevant features R^- to the normalized features $\tilde{F}$, the obtained features $\tilde{F}^-$ should result in a larger entropy as they are less discriminative.

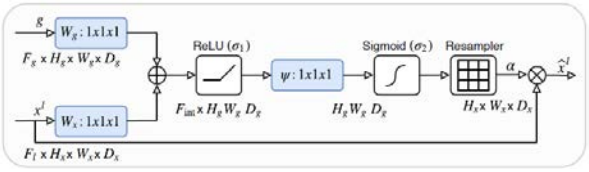


Figure 7: Schematic of additive AG from Oktay et al. (2018).

Regarding the dropout layer, both scenarios with and without the layers were evaluated since the training data might be relatively small for the model to perform well with dropout.

3.4.1.3. ResUNet AG
In this approach, AG (shown in Figure 7) are added with the skip connections in the decoder path to ideally learn to focus on important structures for segmenting MS lesions regardless of the domain.

Due to memory constraints, the ResUNet model was build with 4 layers instead of 5. It was trained with a batch size of 4 and the loss function combined the dice loss and the focal loss. Similarly to the previous approach, both cases with and without dropout layers were evaluated.

3.4.1.4. ResUNet HM
This approach uses the same 3D ResUNet model but trained on data that underwent histogram matching. This simple method consists of transforming the histogram of an input image to match the histogram of another selected image by mapping the voxel intensities accordingly. Ideally, if the model is well trained on the transformed data, it should remain performant with new unseen target data regardless of the domain, as the images will also be transformed with HM. The selected image was chosen from Lyon (MSSEG-1) by observation for having a good contrast, and the training was done with a batch size of 3 and a loss function combining the dice and focal losses.

3.4.2. VAE

3.4.2.1. SegResNetVAE
This model (Myronenko, 2019) is based on an asymmetrical ED architecture, where the encoder is larger than the decoder to extract deep image features. In addition, a VAE branch is added to the encoder endpoint to reconstruct the image. The motivation behind it is to guide and regularize the encoder for a better generalization.

The SegResNetVAE model, shown in Figure 8, is implemented in the MONAI library and is trained with the default configuration (except for the number of output filters of the initial convolutional layer that is 32 instead of 8), a batch size of 4 and a loss function combining the dice, focal and VAE losses.

3.4.2.2. SegResNet
This approach is only examined to determine whether the VAE branch has a positive impact on the model. The architecture and configurations are the same as SegResNetVAE, with only the VAE decoder branch removed, keeping just the segmentation decoder, and the loss function combining dice and focal losses only.

3.4.2.3. ResUNetVAE
In this approach, a ResUNet adapted with a VAE bottleneck and decoder is first trained on the available data to reconstruct the original input images. Here the loss consists of only the VAE loss, and the network output is 1 channel.

Once the model is trained (on a batch size of 3), its encoder is used in evaluation mode to return the bottleneck output and the skip connections, which will be used as inputs to train another decoder (same configuration as ResUNet decoder) for segmenting the lesions.

The goal here is to use the efficiently encoded feature representations from the VAE encoder (ideally close to a standard normal distribution regardless of the data's original domain) as inputs to a segmentation decoder. The loss function for this part was the combination of dice and focal losses, and the batch size is kept as 3.

3.4.3. Transformers

Two additional DL networks based on transformers are trained on a batch size of 1: UNETR (Hatamizadeh et al., 2022b) and SwinUNETR (Hatamizadeh et al., 2022a). Both models are implemented in MONAI and the configurations are kept as default. The loss function used during both trainings was the combination of dice and focal losses.

3.5. Evaluation metrics

The models' performances are evaluated following the common metrics for medical image segmentation between the GT and the predicted segmentation masks. It includes:

- Segmentation and detection wise metrics
  Dice similarity coefficient (DSC), true positive fraction (TPF) and false positive fraction (FPF) are computed both voxel-wise (segmentation) and lesion-wise (detection) as follows:

$$DSC = \frac{2 \times TP}{FN + FP + 2 \times TP}$$

$$TPF = \frac{TP}{TP + FN} \qquad FPF = \frac{FP}{FP + TN}$$

  where TP, TN, FP and FN denote the number of true positives, true negatives, false positives and false negatives, respectively.

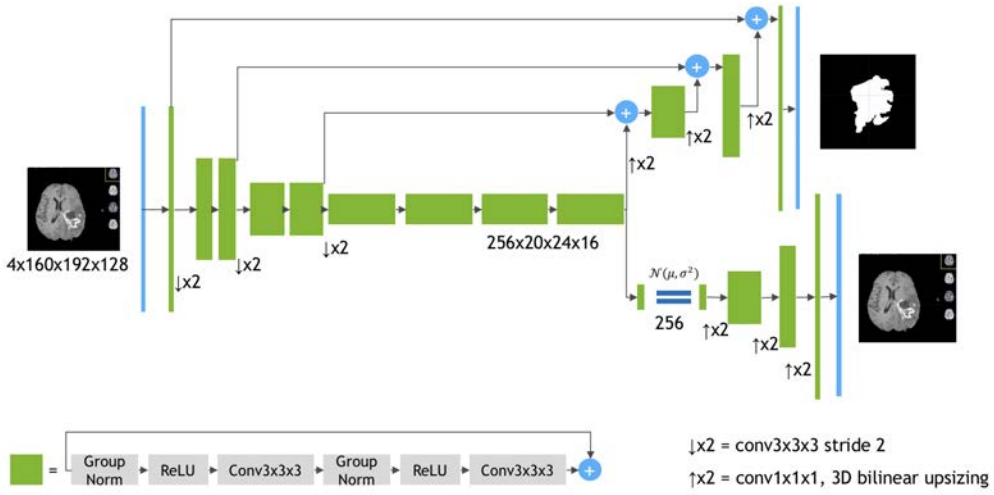


Figure 8: SegResNetVAE architecture from Myronenko (2019) as implemented in MONAI.

- Shifts challenge 2023 metrics
  The Shifts challenge 2023 evaluation was based on two main metrics provided with their baseline (Malinin et al., 2022):
  - Normalized DSC (nDSC): decorrelates model performance and lesion load to avoid having higher DSC just because the patients present larger lesion volumes.
  - Area under the retention curves (nDSC R-AUC): jointly assesses robustness and uncertainty quality from the nDSC scores. The error-retention curve is obtained by replacing at a time portions of the model's predictions by ground-truth labels (in order of decreasing uncertainty) and computing the error metric (here nDSC). The final metric would be the area under the resulting curve.
- Other metrics
  - HD: mesures the Hausdorff distance to estimate how much the predicted segmentation masks deviates from the GT.
  - Abs-lesion-diff: computes the absolute difference between the number of lesions in the GT and the prediction masks.
- Statistical test
  To evaluate the statistical significance of the performance between the different approaches, a serie of permutation tests are ran between the dice scores (DSC_s and DSC_d) of each pair of methods.

  Following the work of Salem et al. (2020), the permutation tests consist of selecting random subsets of images from the dataset (dev_{out}), and for each pair of methods, performing permutations of their DSC values from the selected subsets and counting the number of times that the t-test results in a $p-value \leq 0.05$. This process is repeated S times (here S=1000), and the mean and std of the fraction of times when each method obtained significant *p-values* is calculated over all iterations S. This results in methods with higher significance having higher means. All the approaches are then ranked into three levels according to the mean score ($\mu_0 \pm \sigma_0$) of the best method:
  $rank_1$ has methods with mean scores in (μ_0-σ_0, μ_0], $rank_2$ in (μ_0-$2\sigma_0$, μ_0-σ_0], and $rank_3$ in (μ_0-$3\sigma_0$, μ_0-$2\sigma_0$].

3.6. Implementation details

This project was implemented using Pytorch 1.12.1 and CUDA 11.6 on VSCode IDE. MONAI 0.9.0 was used for its ready-to-use loss functions and data transforms, Tensorboard 2.11.2 for monitoring training plots and ITK-SNAP for visualization. The networks were trained on several Nvidia A30 GPUs devided into MIGs with 12Gb of memory each.

As for the challenge submissions, Docker Engine 20.10.24 was used to build docker images of the best performing methods and upload them on the Grand-Challenge online platform.

4. Results

This section is divided into two main parts. In the first one, we present the results obtained with the different approaches described before and which were trained using the *train* and dev_{in} data from the Shifts challenge 2023. We also present the results of the submitted solutions for the challenge, which were evaluated online on the private dataset of Lausanne. In the second part, we refine the best performing model by adding more images from the in-house dataset in order to include more brain volumes with smaller lesions.

Table 2: The resulting evaluation metrics of the different approaches when tested on the dev_{out} dataset. The methods consist of the reference baseline from the challenge, the ResUNet based, VAE-based and transformers-based approaches, as well as ResUNet SNR trained on both datasets (challenge and in-house) and denoted as 'ResUNet SNR + VH'. The scores represent the mean and std over all patients and are divided into segmentation and detection wise, challenge-specific and other common metrics used for the segmentation task.

	Segmentation			**Detection**			**Challenge**		**Other metrics**	
Model	DSC_s	TPF_s	FPF_s	DSC_d	TPF_d	FPF_d	*nDSC R-AUC*	*nDSC*	*HD*	*Abs-lesion-diff*
UNet (baseline)	53.31 ± 22.26	43.38 ± 21.29	22.71 ± 21.25	29.72 ± 10.51	21.16 ± 8.32	23.75 ± 25.13	2.5389 ± 1.477	51.25 ± 19.42	38.48 ± 11.36	60.79 ± 64.94
ResUNet	69.73 ± 12.69	**65.86 ± 14.51**	24.17 ± 13.38	**52.88 ± 10.53**	37.96 ± 9.47	11.62 ± 12.04	1.1499 ± 0.7571	67.05 ± 8.18	**30.95 ± 9.92**	48.25 ± 50.17
ResUNet SNR	**69.86 ± 13.58**	63.31 ± 16.14	19.41 ± 9.89	52.25 ± 12.15	**40.26 ± 11.56**	11.75 ± 14.14	1.1046 ± 0.7177	66.96 ± 10.39	31.14 ± 10.88	39.33 ± 42.42
ResUNet AG	69.55 ± 15.05	62.17 ± 18.22	**17.41 ± 8.2**	50.05 ± 11.81	36.95 ± 10.86	11.15 ± 13.42	**1.0416 ± 0.6057**	65.76 ± 11.63	32.55 ± 13.26	47.29 ± 48.34
ResUNet HM	69.45 ± 11.84	64.51 ± 11.59	23.5 ± 15.19	52.6 ± 11.66	40.05 ± 10.51	18.63 ± 17.45	1.1252 ± 0.8529	**67.91 ± 6.64**	31.88 ± 10.76	41.88 ± 46.32
SegResNet	63.32 ± 17.24	57.98 ± 18.93	27.91 ± 15.89	42.56 ± 12.24	27.74 ± 9.08	11.65 ± 15.07	1.8532 ± 1.2625	59.47 ± 13.46	37.83 ± 11.91	65.38 ± 64.02
SegResNetVAE	65.07 ± 17.91	62.56 ± 19.54	30.17 ± 17.42	45.73 ± 11.9	29.65 ± 8.86	**10.24 ± 14.72**	2.4728 ± 1.7005	60.41 ± 14.16	35.64 ± 9.27	66.12 ± 65.77
ResUNetVAE	60.82 ± 18.08	52.62 ± 18.1	23.14 ± 15.26	43.48 ± 10.16	34.08 ± 8.39	18.97 ± 16.96	1.6558 ± 0.9717	59.19 ± 15.55	32.47 ± 8.91	41.29 ± 48.26
UNETR	58.03 ± 21.0	50.99 ± 18.36	27.63 ± 24.75	29.59 ± 10.46	29.32 ± 8.81	48.51 ± 24.51	2.3305 ± 1.8897	57.86 ± 15.85	32.07 ± 8.66	40.33 ± 41.57
SwinUNETR	63.05 ± 14.11	53.46 ± 15.82	19.18 ± 11.65	46.38 ± 11.13	39.43 ± 11.78	15.93 ± 13.97	1.8122 ± 1.0336	61.14 ± 10.88	31.18 ± 7.28	**38.0 ± 44.76**
ResUNet SNR + VH	**71.14 ± 15.15**	**67.5 ± 18.09**	22.33 ± 11.34	**55.57 ± 11.78**	**43.27 ± 11.27**	19.04 ± 16.63	1.1279 ± 0.658	**67.35 ± 10.87**	**29.83 ± 9.14**	41.67 ± 44.64

4.1. Training with the challenge dataset only

Different approaches were studied to asses the improvement of the baseline through the different changes and components added to it. In this project, we analyzed the impact on the segmentation results within a domain shift situation of all of: AG, style-normalized relevant features, histogram matching, efficient latent space representations of VAEs and transformers.

The presented results are divided into two subsections: the first one shows the results obtained on the available testing data (dev_{out} and $eval_{in}$), and the second one presents the results obtained on the challenge submissions.

4.1.1. Testing results

The resulting metrics of each of the evaluated methods on the OOD data (dev_{out}) are shown in Table 2, and examples of the obtained segmentation masks are seen in Figure 9. The overall scores seem rather close between the methods and clearly outperform the UNet baseline of the challenge.

The visual analysis of the results depicted in Figure 9 indicates that the majority of the methods managed to segment most of the lesions. However, some methods exhibited a tendency to oversegment certain regions of high intensities (as highlighted by the yellow arrows), while others failed to capture certain lesions (as indicated by the purple arrows). It can also be noted, from this example, that the ResUNet SNR model demonstrated relatively lower levels of oversegmentation compared to the other methods.

Additionally, ResUNet based approaches outperformed the ones relying on VAE and transformers. This observation is further supported by the statistical test conducted for both segmentation and detection DSC (as seen in Figure 10), where ResUNet-based models appear in $rank_1$, followed by the VAE-based and SwinUNETR.

Focusing on the DSC, Figure 11 shows the boxplot of the DSC_s of all methods on both in-domain ($eval_{in}$) and OOD (dev_{out}) data. It can be noticed that the performances on both domains are relatively close for most methods, meaning that the proposed approaches generalize well and achieve good segmentation results, with ResUNet SNR being slightly better. However, in Figure 12, the boxplot of the DSC_d of all methods shows a very large decrease in performance between the two domains, which suggests that all of the trained models are missing a considerable amount of lesions. Nonetheless, ResUNet and ResUNet SNR demonstrated a slightly better performance compared to the others.

To further understand the gaps noticed in the boxplot, we will focus on ResUNet SNR, the best performing model so far in segmentation and detection sensitivity, and inspect some correlation plots. Figure 13 shows on the left side the correlation plot between the true and predicted lesion volumes in the OOD data. We can note that the overall predicted volumes tend to be less than the real ones, but not in very large proportions, which corresponds to the relatively good segmentation scores obtained. As for the right side of Figure 13, the correlation plot between the real and predicted number of lesions in the OOD data shows that even the best per-

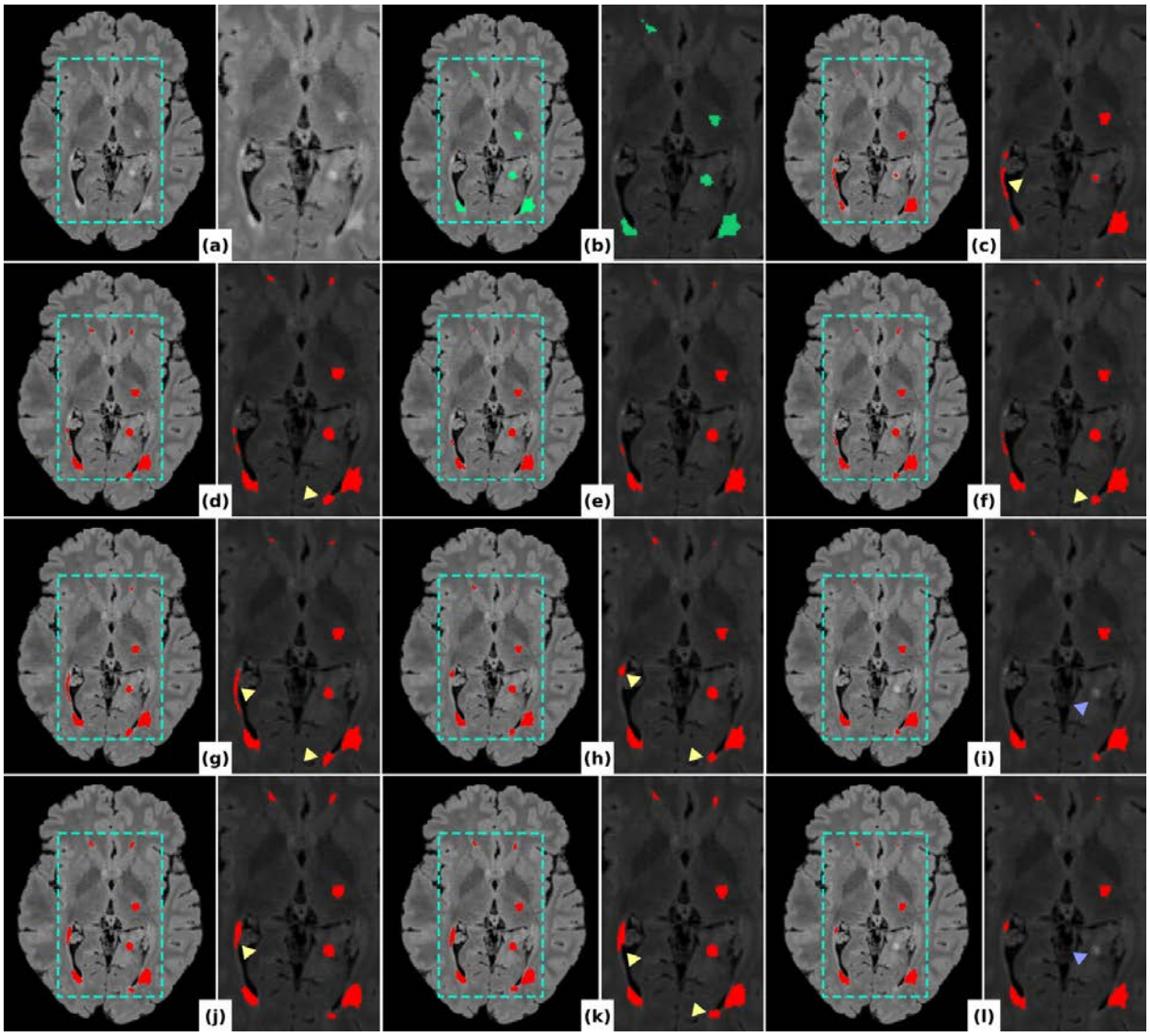


Figure 9: Qualitative results for MS lesion segmentation. The right side of each pair of images depicts zoomed regions of interests shown in blue rectangles on whole-brain scan (on the left side). (a) and (b) represent the FLAIR and GT respectively, while the remaining pairs represent the overlayed segmentation masks obtained from each method: Baseline UNet (c), ResUNet (d), ResUNet SNR (e), ResUNet AG (f), ResUNet HM (g), UNETR (h), SwinUNETR (i), SegResNet (j), SegResNetVAE (k), and ResUNetVAE (l). The yellow arrows indicate regions of oversegmentation and the purple ones indicate missed lesions.

forming model is missing a large number of lesions, almost by a factor of 2, yielding in low metric values. This might be due to the training data not having enough small lesions to train the models on, especially that the dataset is limited and the 3D patches are mostly centered in lesions.

For this reason, the next step was to further train the best model (ResUNet SNR) on some more images from the in-house dataset and evaluate its impact on the results.

4.1.2. Challenge submission results

Due to the limited time that was left before the submission deadline, only two of the submitted approaches were based on the methods proposed in this study, which are ResUNet and ResUNet SNR.

Table 3 shows the scores obtained on the private $eval_{out}$ dataset used for evaluation (Lausanne dataset). The ranking of the solutions was based on the nDSC R-AUC scores, and the total number of submissions that appear on the leaderboard was 36. The ResUNet model achieved the best nDSC score among all participants, but was ranked 8^{th} according to the nDSC R-AUC metric. As for the ResUNet SNR method, it reached the 6^{th} position on the leaderboard, making our team in the 4^{th} place.

4.2. Training with challenge and in-house datasets

Considering the encountered issue in the previous results (subsection 4.1.1), more data was added to train

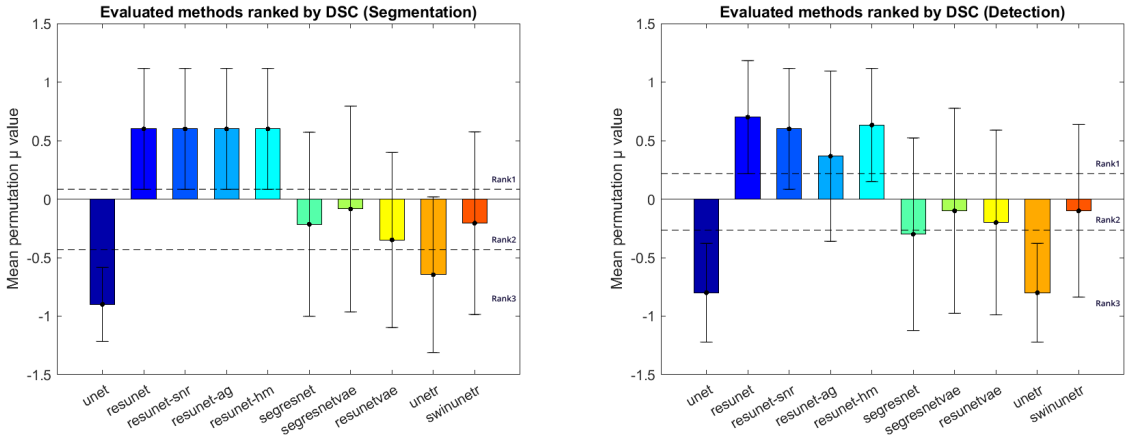


Figure 10: Permutation test on dice scores (DSC_s and DSC_d).

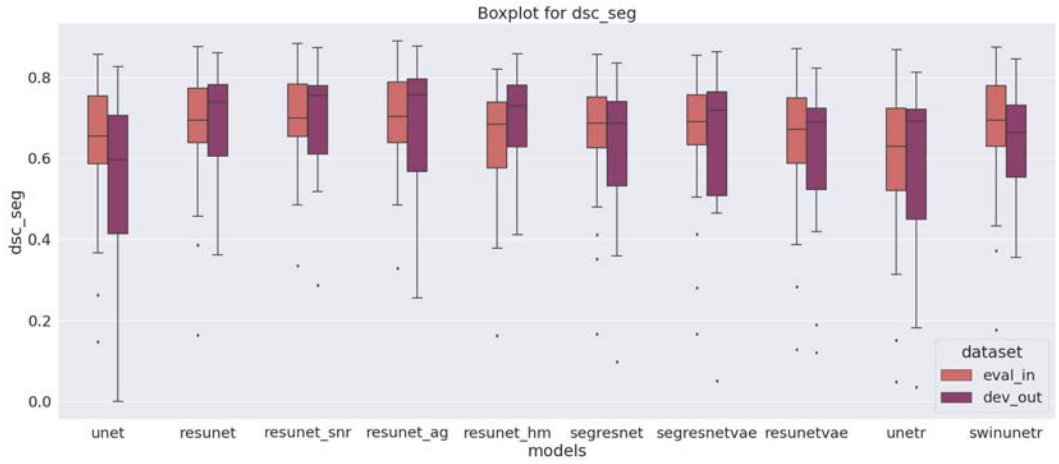


Figure 11: Boxplots of DSC_s using the different approaches.

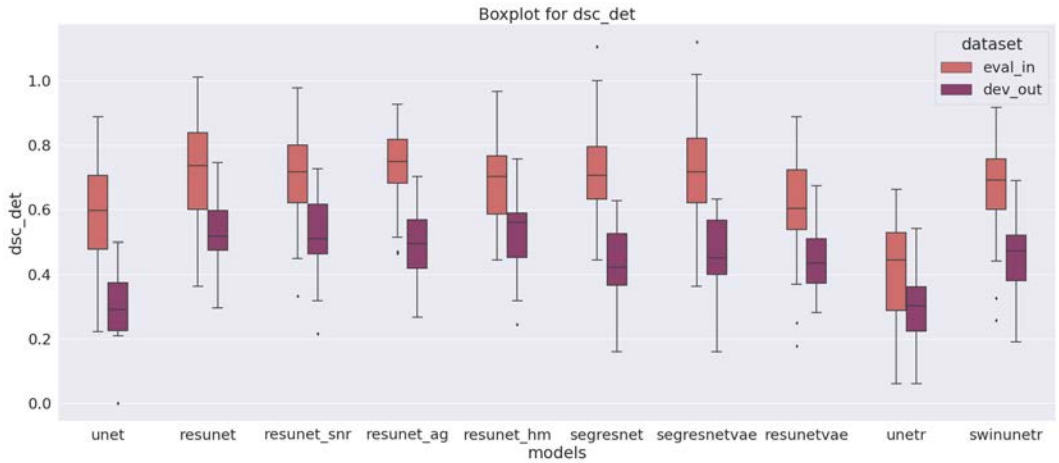


Figure 12: Boxplot of DSC_d using the different approaches.

the previous ResUNet SNR model. Table 4 shows the average lesion volume for each data partition. It can be seen that the *train* set has some lesions of very large sizes, while the dev_{out} seems to have a larger number of small lesions. As for the VH data, both partitions have a small lesion tendency, mainly in the *VH train*, which led us to add it with the *train* data of the challenge to further tune the model.

To achieve that, the ResUNet SNR model is first loaded with the previously trained weights, then trained

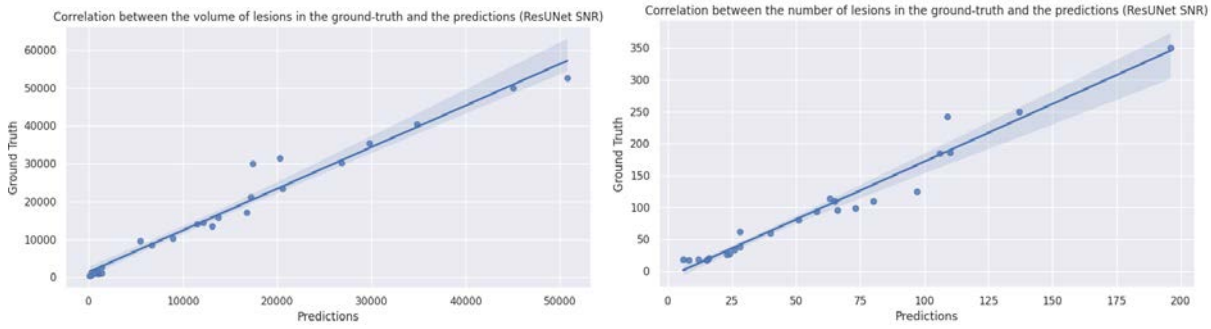


Figure 13: Correlation plots using ResUNet SNR model. From left to right: correlation plot between the true and predicted lesion volumes, and correlation plot between the true and predicted lesion numbers.

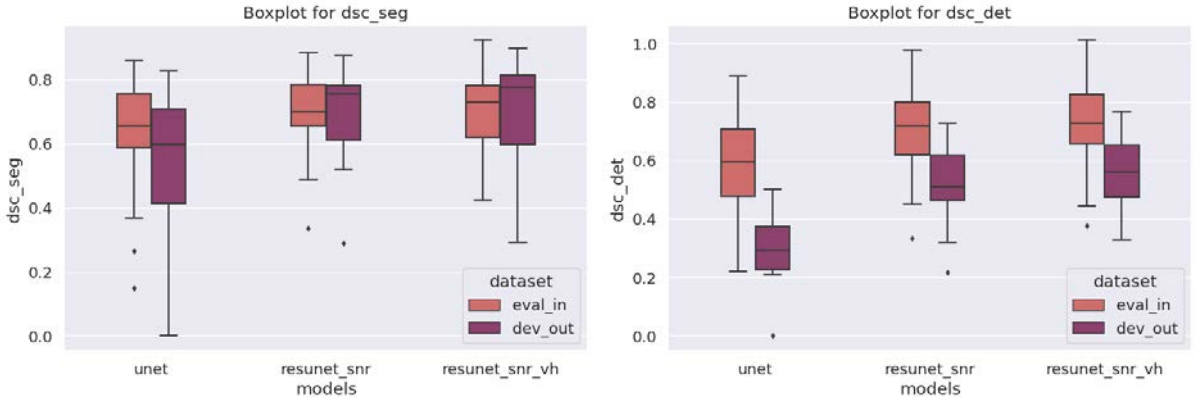


Figure 14: Boxplot of DSC_s and DSC_d (from left to right, respectively) using ResUNet SNR with the combined training data.

Table 3: Submission scores of the two proposed solutions for the Shifts challenge 2023, along with the current winner's scores for a reference.

Model	nDSC R-AUC	nDSC	Rank / 36
Current winner	1.28 ± 1.69	51.10	1
ResUNet SNR	1.60 ± 1.24	60.04	6
ResUNet	1.71 ± 1.83	66.87	8

Table 4: Additional characteristics of the datasets, in terms of total number of lesions and their average volumes.

	Train	Dev_{out}	*VH train*	$VH\ eval_{in}$
Total lesion count	1628	3544	790	573
average lesion volume (mm^3)	376 ± 2994	120 ± 1254	156 ± 427	183 ± 564

on the new combined data with dev_{in} for validation, a batch size of 4 and a starting LR of $1e^{-4}$. Even though the preprocessing of the two datasets are not the same, the little differences that might appear in the in-house dataset can be considered as part of its own domain.

The obtained results seen in Table 2 show a little improvement in terms of DSC and sensitivity (TPF). The comparison between both cases and the baseline is also seen in Figure 14. However, a t-test between the DSC before and after adding the data gave a p-value of 0.97 and 0.85 in segmentation and detection respectively, which implies that the results after adding the images have low statistical significance. This suggests that further model improvements are still necessary to enhance the segmentation performances.

5. Discussion

5.1. MS lesion segmentation

In this project, we investigated some methods and architectures that would potentially help surpass the domain shift problem encountered when segmenting MS lesions from MRI scans. In this subsection, we try to explain the possible reasons behind the performances of each approach.

From the different quantitative results, ResUNet achieved higher scores than the baseline UNet, thanks to the residual units that allow a better flow of information in the network, as well as the use of IN, which helps style-normalizing features and making them more generalizable.

By adding the SNR module in the encoder part and removing the dropout layers, the ResUNet SNR approach achieved the best metrics among all methods, in terms of segmentation and detection (sensitivity), and had the second lowest absolute lesion difference score. This reinforces the idea that restituing task-relevant feature representations would help better generalize the

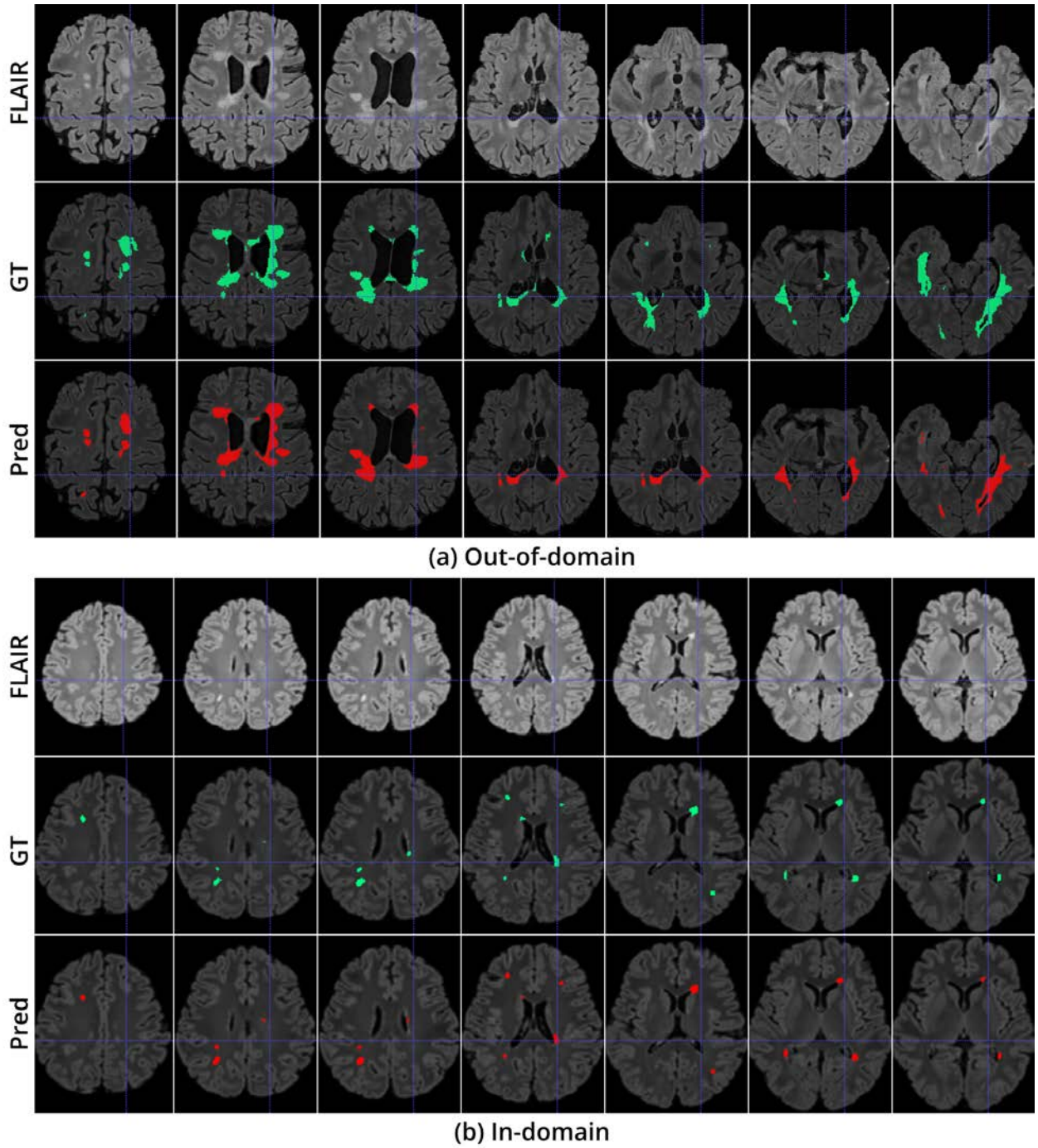


Figure 15: Qualitative results for both in-domain and out-of-domain cases, showing segmentation outputs of different slices using the ResUNet SNR model. Both cases were among the ones achieving the best metrics.

model across domains, especially when using the dual restitution loss function that helped obtaining enhanced features. More examples of the resulting segmentation masks for this approach are shown in Figure 15 and Figure 16. The first one presents different slices from two of the best segmented cases (one for each domain). In the OOD case, the lesion load was big, and the model managed to detect most of the lesions and have a good overall segmentation. As for the in-domain case, the absence of domain shift allowed even smaller lesions to be segmented and detected. However, the model failed in other cases, as seen in Figure 16. Here the model missed small lesions in both in- and out-of-domain cases, highlighting the necessity for additional optimization of the model

Another added module in the ResUNet is the AG. From Figure 17, it can be seen that the attention mechanism helped the model (without dropout) focus its fea-

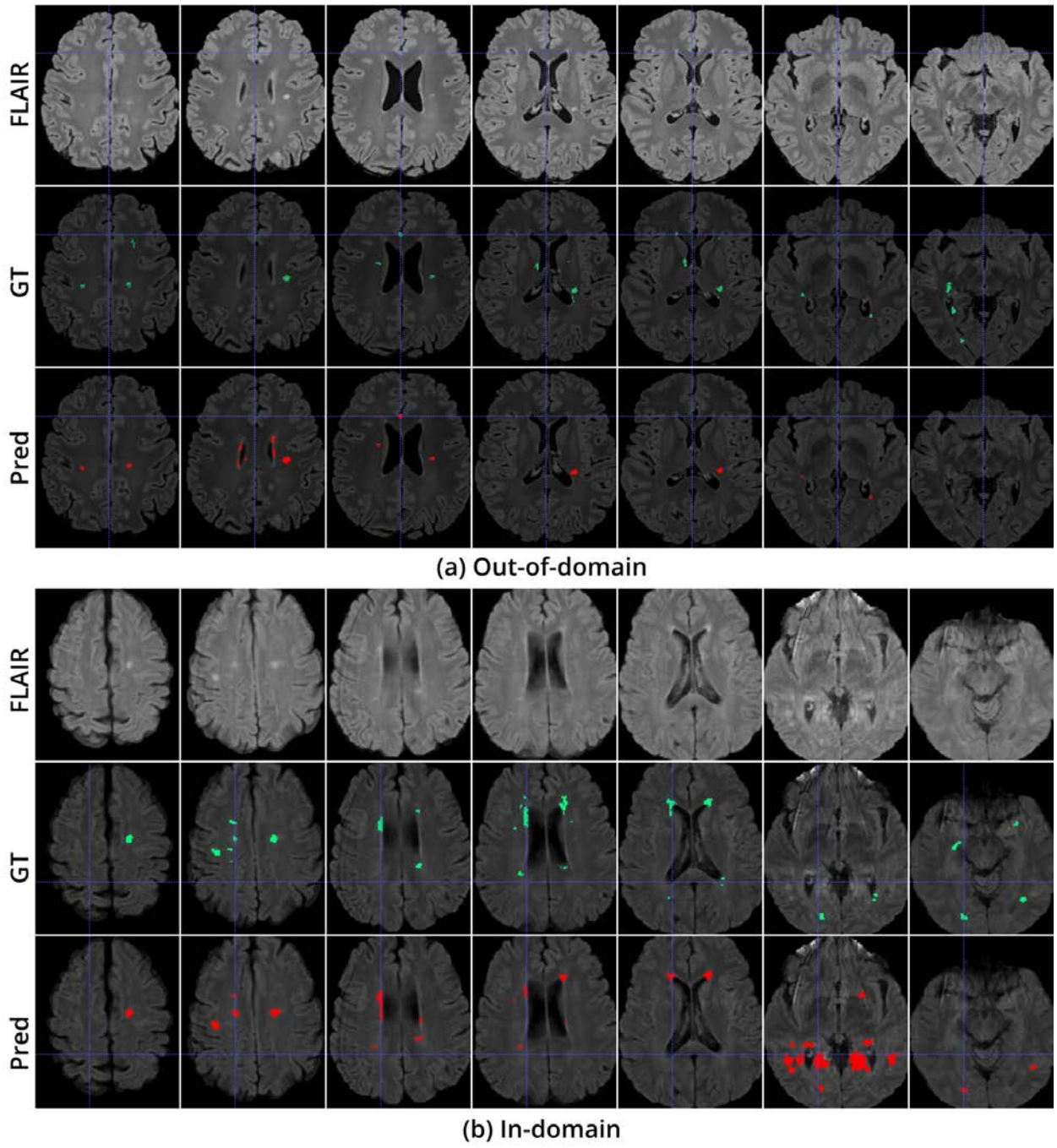


Figure 16: Qualitative results for both in-domain and out-of-domain cases, showing segmentation outputs of different slices using the ResUNet SNR model. Both cases were among the ones getting the lowest metrics.

tures a little more on the main important regions of the images, corresponding to lesions. Even though the overall obtained results are lower than the ResUNet, this approach still managed to achieve the lowest FPF scores, making it among the most reliable ones. Its robustness can also be seen through the challenge metric (nDSC R-AUC) in Table 2. This suggests that the model, if further optimized and added with SNR module, might be more fit to tackle the small lesions issue while facing the domain shift.

Considering the inference-time constraint of the challenge, histogram-matching method was tested with the ResUNet model. It achieved comparable results with the others, which further confirms the simplicity and effectiveness of HM. However, this approach is still prone to unstability as it relies on the selected image for HM, and thus might fail in some other cases.

Moving on to the VAE based models, comparing the results between SegResNet and SegResNetVAE shows a little improvement when the VAE branch is added.

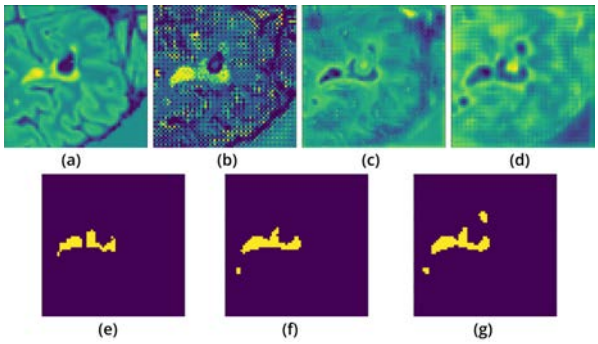


Figure 17: Comparison between segmentation masks with and without attention. (a) input patch, (b) attention map, (c) feature map with attention, (d) feature map without attention, (e) GT, (f) segmentation mask with attention, and (g) segmentation mask without attention.

This confirms that the VAE branch allows to regularize the encoded features and hence, helps achieving better performances. However, it is not as performant as ResUNet, which might be due to the default normalization layer being the group norm instead of IN.

As for ResUNetVAE, even though it achieved a better sensitivity score (detection), the DSC are lower compared to SegResNetVAE. This might point out the importance of training both models (reconstruction and segmentation) jointly for a better regularization of the encoded features towards a good semantic segmentation.

Regarding the transformers-based approaches, the underperformance of the UNETR might be explained by the complexity of the network and its need for a larger amount of data to train on. In addition, the size of the input images are rather small patches, while the transformer is supposed to be efficient for capturing global multi-scale information. This information might be lost when using patches as the transformer will further divide them into tokens.

Nonetheless, SwinUNETR still managed to outperform UNETR, probably thanks to the efficiency of the SwinTransformers for focusing on nearby patches and capturing local context within each input patch.

At the end, the different approaches performed quite similar to each others, with the ResUNet based ones leading in most metrics and having higher statistical significance. In a DG point of view, the ResUNet SNR is the model that mainly tackles the domain shift problem, and its limited efficiency could be linked to the training data. In fact, this hypothesis can be supported by the results obtained when testing the model trained on the challenge data directly on the VH dataset. The DSC scores dropped to the values of 37.37 ± 21.51 and 44.71 ± 20.91 in segmentation and detection, respectively. The VH dataset contains smaller lesion loads, and thus the model did not manage to generalize well. Also, when more data from the VH dataset was added for training, the results increased a little bit, as more examples of small lesions were provided to train the model on.

With that being said, there is still a clear need for robust models for small lesions that are more frequent in MS, as well as the need to highlight the detection metrics and sensitivity, which are far more important for doctors in this kind of diseases.

5.2. Limitations and future work

It is clear that the easy way to achieve the best domain generalization is to include all possible variations of the data when training a model. As it is not practically feasible with the scarcity of medical data and the lack of representative data that capture the variabilities across domains and MRI scanners, we can only bypass these limitations by working on improving the models with the available data in hand.

From this project, we can note that the heterogeneity of the disease, in terms of lesion variations in shapes and sizes, considerably affects the results and must be tackled jointly with the domain shift problem. For this, a better optimization of the proposed models is vital, by tuning the hyperparameters and including more adequate data transforms and augmentation. This latter can be based on the MixUp strategy and is kept for a future study. We also plan to make use of pre-trained models from MONAI and include more modalities in training for a more robust solution, as well as exploring other feature disentanglement methods to further highlight the content part and normalize the style ones.

6. Conclusions

In conclusion, several approaches have been explored, taking advantage of some state-of-the-art strategies in DG. The methods included architectures based on ResUNet, VAE and transformers. According to the results, it appears that residual UNets are still favorable for the imposed constraints, and perform even better when added with the SNR module. The effectiveness of its relevant feature learning allowed to capture the essential content of the images, yielding in a relatively good segmentation. However, it is important to note that there is still need for improvement in the lesion detection capability of the model, mainly in detecting smaller lesions that were frequently missed.

Nonetheless, during the Shifts challenge 2023, the ResUNet SNR model still proved its efficiency and made it among the top solutions in the leaderboard, allowing the team to reach the 4^{th} position among all participants.

At the end, despite the presence of domain shifted data, MS lesion segmentation remains a challenging task due to the inherent heterogeneity of the disease and the lack of representative data that encapsulates these

variabilities. Therefore, more robust architectures and optimization techniques must be explored to surpass the current results.

Acknowledgments

I would like to thank my supervisors Dr Xavier Lladó and Dr Arnau Oliver for providing me with valuable support, guidance and feedback throughout this master thesis, and for the opportunity to work on it within the ViCOROB institute and exchange with its PhD students.

I am also truly grateful to my family and friends for their continuous support and encouragements, as well as my MaIA classmates for this amazing journey.

This work was carried out in collaboration with The Observatoire Français de la Sclérose en Plaques (OFSEP), who is supported by a grant provided by the French State and handled by the"Agence Nationale de la Recherche," within the framework of the "Investments for the Future" program, under the reference ANR-10-COHO-002, by the Eugène Devic EDMUS Foundation against multiple sclerosis and by the ARSEP Foundation.

References

Ackaouy, A., Courty, N., Vallée, E., Commowick, O., Barillot, C., Galassi, F., 2020. Unsupervised domain adaptation with optimal transport in multi-site segmentation of multiple sclerosis lesions from mri data. Frontiers in computational neuroscience 14, 19.

Billast, M., Meyer, M.I., Sima, D.M., Robben, D., 2020. Improved inter-scanner ms lesion segmentation by adversarial training on longitudinal data, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I, Springer. pp. 98–107.

Carass, A., Roy, S., Jog, A., Cuzzocreo, J.L., Magrath, E., Gherman, A., Button, J., Nguyen, J., Prados, F., Sudre, C.H., et al., 2017. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. NeuroImage 148, 77–102.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3d u-net: learning dense volumetric segmentation from sparse annotation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19, Springer. pp. 424–432.

Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., Pop, S.C., Girard, P., Ameli, R., Ferré, J.C., et al., 2018. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. Scientific reports 8, 13650.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. The journal of machine learning research 17, 2096–2030.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D., 2022a. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I, Springer. pp. 272–284.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D., 2022b. Unetr: Transformers for 3d medical image segmentation, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 574–584.

Hu, S., Liao, Z., Zhang, J., Xia, Y., 2022. Domain and content adaptive convolution based multi-source domain generalization for medical image segmentation. IEEE Transactions on Medical Imaging 42, 233–244.

Ilse, M., Tomczak, J.M., Louizos, C., Welling, M., 2020. Diva: Domain invariant variational autoencoders, in: Medical Imaging with Deep Learning, PMLR. pp. 322–348.

Jin, X., Lan, C., Zeng, W., Chen, Z., 2021. Style normalization and restitution for domain generalization and adaptation. IEEE Transactions on Multimedia 24, 3636–3651.

Kamraoui, R.A., Ta, V.T., Tourdias, T., Mansencal, B., Manjon, J.V., Coup, P., 2022. Deeplesionbrain: Towards a broader deep-learning generalization for multiple sclerosis lesion segmentation. Medical Image Analysis 76, 102312.

Kerfoot, E., Clough, J., Oksuz, I., Lee, J., King, A.P., Schnabel, J.A., 2019. Left-ventricle quantification using residual u-net, in: Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges: 9th International Workshop, STACOM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers 9, Springer. pp. 371–380.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 .

Li, D., Yang, Y., Song, Y.Z., Hospedales, T., 2018a. Learning to generalize: Meta-learning for domain generalization, in: Proceedings of the AAAI conference on artificial intelligence.

Li, H., Pan, S.J., Wang, S., Kot, A.C., 2018b. Domain generalization with adversarial feature learning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5400–5409.

Li, K., Kong, L., Zhang, Y., 2020. 3d u-net brain tumor segmentation using vae skip connection, in: 2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC), IEEE. pp. 97–101.

Li, L., Gao, K., Cao, J., Huang, Z., Weng, Y., Mi, X., Yu, Z., Li, X., Xia, B., 2021a. Progressive domain expansion network for single domain generalization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 224–233.

Li, L., Zimmer, V.A., Schnabel, J.A., Zhuang, X., 2021b. Atrialgeneral: Domain generalization for left atrial segmentation of multicenter lge mris, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24, Springer. pp. 557–566.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988.

Lladó, X., Oliver, A., Cabezas, M., Freixenet, J., Vilanova, J.C., Quiles, A., Valls, L., Ramió-Torrentà, L., Rovira, À., 2012. Segmentation of multiple sclerosis lesions in brain mri: a review of automated approaches. Information Sciences 186, 164–185.

Malinin, A., Athanasopoulos, A., Barakovic, M., Cuadra, M.B., Gales, M.J., Granziera, C., Graziani, M., Kartashev, N., Kyriakopoulos, K., Lu, P.J., et al., 2022. Shifts 2.0: Extending the dataset of real distributional shifts. arXiv preprint arXiv:2206.15407 .

Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 fourth international conference on 3D vision (3DV), Ieee. pp. 565–571.

Myronenko, A., 2019. 3d mri brain tumor segmentation using autoencoder regularization, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop,

BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4, Springer. pp. 311–320.

Nam, H., Kim, H.E., 2018. Batch-instance normalization for adaptively style-invariant neural networks. Advances in Neural Information Processing Systems 31.

Nam, H., Lee, H., Park, J., Yoon, W., Yoo, D., 2021. Reducing domain gap by reducing style bias, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8690–8699.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 .

Pan, X., Luo, P., Shi, J., Tang, X., 2018. Two at once: Enhancing learning and generalization capacities via ibn-net, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 464–479.

Popescu, B.F.G., Pirko, I., Lucchinetti, C.F., 2013. Pathology of multiple sclerosis: where do we stand? CONTINUUM: Lifelong Learning in Neurology 19, 901.

Salem, M., Ahmed Ryan, M., Oliver i Malagelada, A., Hussain, K.F., Lladó Bardera, X., 2022. Improving the detection of new lesions in multiple sclerosis with a cascaded 3d fully convolutional neural network approach. Frontiers in Neuroscience, 2022, vol. 16, art. núm. 1007619 .

Salem, M., et al., 2020. Deep learning methods for automated detection of new multiple sclerosis lesions in longitudinal magnetic resonance images .

Shi, Y., Seely, J., Torr, P.H., Siddharth, N., Hannun, A., Usunier, N., Synnaeve, G., 2021. Gradient matching for domain generalization. arXiv preprint arXiv:2104.09937 .

Somavarapu, N., Ma, C.Y., Kira, Z., 2020. Frustratingly simple domain generalization via image stylization. arXiv preprint arXiv:2006.11207 .

Valverde, S., Salem, M., Cabezas, M., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Rovira, À., Salvi, J., Oliver, A., Lladó, X., 2019. One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. NeuroImage: Clinical 21, 101638.

Walton, C., King, R., Rechtman, L., Kaye, W., Leray, E., Marrie, R.A., Robertson, N., La Rocca, N., Uitdehaag, B., van Der Mei, I., et al., 2020. Rising prevalence of multiple sclerosis worldwide: Insights from the atlas of ms. Multiple Sclerosis Journal 26, 1816–1821.

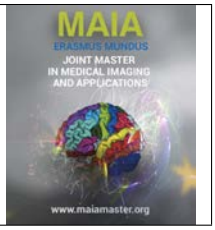
Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., Yu, P., 2022. Generalizing to unseen domains: A survey on domain generalization. IEEE Transactions on Knowledge and Data Engineering .

Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2017. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 .

Zhou, K., Yang, Y., Qiao, Y., Xiang, T., 2021. Domain adaptive ensemble learning. IEEE Transactions on Image Processing 30, 8008–8018.

Medical Imaging and Applications

Master Thesis, June 2023



Breast Mass Detection and Classification using Transfer Learning on OPTIMAM Dataset through RadImageNet weights

Ruth Kehali Kassahun, Mario Molinara

Department of Electrical and Information Engineering, University of Cassino and Southern Lazio, Cassino, Italy

Abstract

A significant number of women are diagnosed with breast cancer each year. Early detection of breast masses is crucial in improving patient prognosis and survival rates. In recent years, deep learning techniques, particularly object detection models, have shown remarkable success in medical imaging, providing promising tools for the early detection of breast masses. This thesis uses transfer learning methodologies to present an end-to-end breast mass detection and classification pipeline. Our approach involves a two-step process: initial detection of breast masses using variants of the You Only Look Once (YOLO) object detection models, followed by classification of the detected masses into benign or malignant categories. We used a subset of OPTIMAM (OMI-DB) dataset for our study. We leveraged the weights of RadImageNet, a set of models specifically trained on medical images, to enhance our object detection models. Among the publicly available RadImageNet weights, DenseNet-121 coupled with the yolov5m model gives 0.718 mAP at 0.5 IoU threshold and a True Positive Rate (TPR) of 0.97 at 0.85 False Positives Per Image (FPPI). For the classification task, we implement a transfer learning approach with fine-tuning, demonstrating the ability to classify breast masses into benign and malignant categories effectively. We used a combination of class weighting and weight decay methods to tackle the class imbalance problem for the classification task.

Keywords: , Breast Mass Detection, Breast Mass Classification, Breast Cancer, RadImageNet, YOLO Object Detection, Transfer Learning, Computer Aided Diagnosis

1. Introduction

Breast cancer is a significant public health concern and one of the most prevalent cancers affecting women predominantly. It is the second leading cause of cancer death among women, following lung and bronchus cancer. (Mattiuzzi and Lippi, 2019) In the year 2022, it attributed to 31% of all women's cancers. Despite advances in screening and treatment, breast cancer remains challenging to diagnose, making early detection and prevention critical to reducing mortality rates. (Siegel et al., 2023)

One of the primary methods for detecting breast cancer is through mammography screening, which involves taking X-ray images of the breast to visualize the internal structure of the breast. Early mammography screening for breast cancer is a widely used clinical procedure. By enabling the detection of breast cancer in its initial stages, when it is more manageable, and the chances for effective treatment are higher, mammography screenings significantly contribute to reducing breast cancer mortality rates.(Tabár et al., 2018). Full field digital mammography (FFDM) clearly depicts low-contrast lesions in dense breasts.

Breast masses, also known as breast lumps, are swellings, bulges, or bumps in the breast that differ from the surrounding tissue. They vary significantly in size, shape, and texture. These masses, such as fibroadenomas or cysts, can be benign (non-cancerous), and some can be malignant (cancerous). Regarding their relationship with breast cancer, not all breast masses indicate malignancy. However, they are often the first noticeable symptom of breast cancer. Therefore, any new or unusual breast mass suggests medical evaluations.

Figure 1 shows that breast masses occasionally present themselves in a way that distinctly separates them from the surrounding breast tissue, enabling a

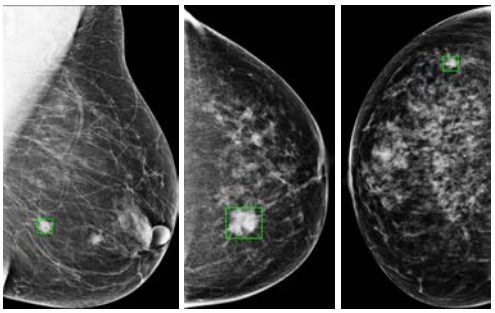


Figure 1: Examples of some breast masses visualized in mammographic images within a distinctive bounding box.

trained radiologist to identify these abnormalities easily. However, the detection of breast masses is not always a straightforward task. At times, the anomalies may be subtly embedded within the breast tissue, making it challenging for even experienced radiologists to distinguish them (Evans et al., 2016). Additionally, the variety in the shape, size, and density of breast masses further complicates their identification. Some masses may appear as tiny specks or calcifications, while others might present as larger, irregularly shaped structures. (Sampat et al., 2005) The characteristics of these masses can be highly variable. This heterogeneity in presentation often makes it difficult to consistently detect breast masses, emphasizing the need for advanced imaging techniques and tools to aid in accurate diagnosis.

This thesis proposes an approach to breast cancer diagnosis by developing a computer-aided diagnosis (CAD) system. This system aims to detect and classify masses present in mammography images from the OPTIMAM (OMI-DB) dataset. To accomplish this, state-of-the-art object detection algorithms, particularly YOLO(You Only Look Once) based algorithms are utilized to detect and localize breast masses accurately.

In addition, the thesis leverages the RadImageNet-trained weights. The research uses weights initialized by the MS COCO dataset along with RadImageNet dataset weights. MS COCO is an object detection and captioning dataset known for its high-quality annotations and diversity of object categories. It is an ideal choice for weight initialization in object detection tasks (Lin et al., 2014) . Alongside, the RadImageNet dataset comprises five million professionally annotated medical images for effective transfer learning (Mei et al., 2022). Consequently, the detected masses are further classified into a respective benign and malignant class as it is crucial for making informed decisions regarding patient treatment plans.

2. State of the art

Current clinical methods for breast mass detection are largely based on radiologist interpretation of mammographic images. Radiologists use various indicating factors such as the shape, margin, and density of abnormal tissue to determine the presence of mass (Sechopoulos et al., 2021). In some cases, computer-aided diagnosis (CAD) systems were used. These systems are designed to assist radiologists in identifying suspicious areas within mammograms that could represent masses, calcifications, or other abnormalities of breast cancer. These systems are practical to reduce oversight errors, typically for inexperienced radiologists. While some studies supported the benefits of CAD in enhancing cancer detection, others raised concerns about increased false-positive rates (Zahoor et al., 2020).

The rise of deep learning, machine learning, and artificial neural networks has greatly contributed to the effective implementation of CAD systems. In the past few years, deep learning models have been increasingly used to detect and classify breast masses (Rodriguez-Ruiz et al., 2019). Various deep learning architectures, mainly convolutional neural networks (CNNs) and, in recent times, transformer-based models, are now being used for breast mass detection and classification tasks. Current research in this area is focused on improving the accuracy of these deep learning models and integrating them into clinical workflows. Transfer learning, where models pre-trained on large, diverse datasets are fine-tuned on specific tasks, is being actively explored to leverage the power of deep learning even when medical imaging datasets are relatively small. (Shin et al., 2016)

For breast mass detection, various CNN-based object detection algorithms have been proposed. (Akselrod-Ballin et al., 2019) proposes Faster R-CNN model to detect breast masses by classifying the dataset into benign, malignant, and other categories. The model achieves an area under the curve (AUC) of 0.91 (95% CI: 0.89, 0.93), with a specificity of 77.3% (95% CI: 69.2%, 85.4%) at a sensitivity of 87%.

Convolutional Neural Networks (CNNs) have been extensively employed to classify breast masses, particularly in the Digital Database for Screening Mammography (DDSM) dataset (Lévy and Jain, 2016) applied CNNs to classify breast masses in the DDSM dataset using different CNN architectures, specifically shallow CNN, AlexNet, and GoogLeNet.

(Yan et al., 2021) used You-Only-Look-Once (YOLO) region proposals for effective detection of breast masses in INbreast and DDSM-CBIS (Digital Database for Screening Mammography) datasets using both patch level and dual view mammographs. In this study, they integrated the mass matching technique and achieved 94.78% as Area Under the Curve(AUC) score for detection and a classification accuracy of 0.87.

(Agarwal et al., 2020) paper used a subset of OMI-DB dataset and applied Faster-RCNN model in a in Full-Field Digital Mammograms (FFDM). In this study, the detection model was tested on INbreast dataset. In

the following dataset it was managed to achieve a true positive rate of 0.87 at 0.84 false positive per image (FPPI). This study serves as one of first research to benchmark on large-scale OPTIMAM Mammography Image Database (OMI-DB).

(Betancourt Tarifa et al., 2023) showcases the potential of transformer models, when combined with convolutional layers for prediction tasks, to achieve remarkable results. This study uses multi scale swin transformers as a backbone model along with Representative Points and the Deformable Detection Transformer (DETR). This research notably achieved a high TPR of 0.903 at 0.8 FPPI, demonstrating the efficacy of the proposed method. It importantly underlines the potential of transformer models, when combined with convolutional layers for prediction tasks, to achieve remarkable results. Providing a strong foundation for future research to explore and leverage the full potential of transformers in the field of medical image analysis.

The work presented by (Ryspayeva and Molinara, 2022) proposes a two-stage methodology for the detection and classification of breast masses in OPTIMAM (OMI-DB) dataset. For breast mass detection, the author used RetinaNet variation of ResNet backbones alongside different weight initialization mainly ImageNet ,COCO weights and model trained from scratch. Additionally, the study emphasized not only on analyzing the whole mammograph but also patches taken from the mammograms images. The results shows a True Positive Rate (TPR) of 0.959 at 0.84 False Positives Per Image (FPPI) when using the RetinaNet with the ResNet151 backbone and ImageNet weights.

3. Material and methods

3.1. Dataset

The OPTIMAM Mammography Image Database (OMI-DB) is a collection of mammography images. The database The collection includes digitized mammograms that have been gathered from various United Kingdom(UK) hospitals and clinics. It includes full annotations for each image, including the radiologist's notes, patient information, and biopsy results. Currently, there are 2.5 million images gathered from 173,319 women from three main UK breast screening centers. OMI-DB dataset may vary based on the specific access agreement and version. Generally it includes digital mammograms captured from the mediolateral oblique (MLO) and craniocaudal (CC) views (Halling-Brown et al., 2021).

For this particular study, we have a total of 7,629 images. 3,529 of the mammograms are identified as having breast masses, and the remaining 4,100 mammograms have no breast masses.

For the breast mass detection task only the cropped region with the breast area is used, as it can be seen in the Figure 3. The dataset consists both left and right mediolateral oblique (MLO) and craniocaudal(CC) views.

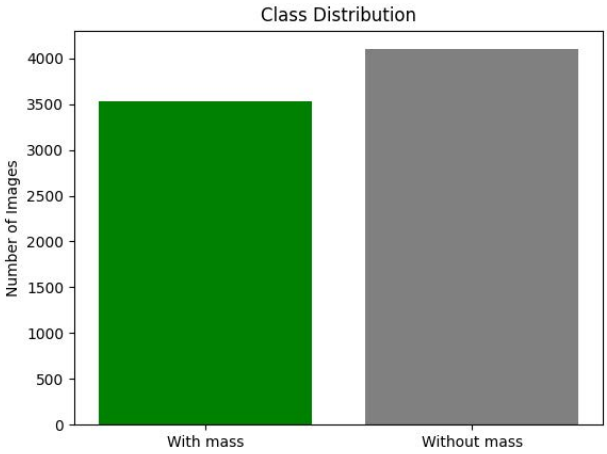


Figure 2: Class Distribution in the dataset

Each image with a breast mass has a corresponding CSV file containing the ground truth annotation for the region of the mass. This annotation is provided in the form of coordinates representing the top-left and bottom-right points of the mass region, denoted as (x1, y1, x2, y2). These coordinates serve to draw the boundaries of the identified breast mass, allowing for precise localization and further analysis. By associating each image with its respective CSV file, the dataset provides essential information for accurately interpreting and understanding the characteristics of breast masses in the images.

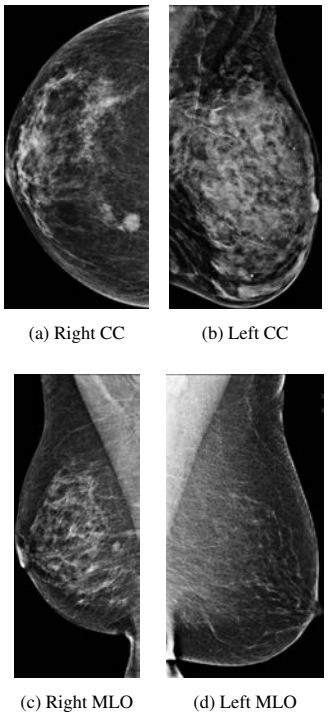
3.2. Bounding Box Conversion

At the beginning of our study, we were provided with ground truth bounding boxes formatted in accordance with the Pascal VOC dataset. The Pascal VOC format describes the bounding box location using four coordinates: the minimum x (x min) and y (y min) values and the maximum x (x max) and y (y max) values. This format essentially provides the top-left and bottom-right corners of the bounding box.

$$[x\_\,\text{min}, y\_\,\text{min}, x\_\,\text{max}, y\_\,\text{max}]$$

In the YOLO format, each bounding box is described using four different parameters: the x and y coordinates of the box's center (x_center and y_center), and the box's width and height. In the YOLO format, the x_center and y_center values represent normalized coordinates, indicating the central point of the bounding box rather than its corners.

$$[\text{ x\_center}, y\_\text{center}, \text{width}, \text{height}]$$



(a) Right CC (b) Left CC

(c) Right MLO (d) Left MLO

Figure 3: Mediolateral oblique (MLO) and Craniocaudal(CC) views in the dataset

We processed each positive case image individually to convert these bounding boxes to the required YOLO format.

3.3. Proposed Method

In this study, our objective is to develop an effective and robust algorithm for detecting breast masses, focusing on utilizing YOLO (You Only Look Once) based object detection techniques. YOLO has emerged as a popular and powerful approach in the field of computer vision due to its real-time performance and accurate object localization capabilities. (Al-Masni et al., 2018)

To achieve our goal, we explored and implemented various versions of the YOLO algorithm, including YOLOv5, YOLOv6, YOLOv7, and YOLOv8. Each version offers distinct architectural enhancements and optimization strategies, which we carefully evaluated and compared in our experiments. Our evaluation consists of various metrics, such as mean average precision(mAP), precision, recall, and true positive rate(TPR) per false positive per image(FPPI), to ensure the assessment of each algorithm's performance.

Our proposed method involves a two-step procedure: breast mass detection followed by a classification task. The object detection stage is mainly uses variants of YOLO models. As shown in Figure 4, the first step involves feeding the mammogram images into the YOLO model. This model identifies the regions in the images containing potential breast masses. A typical YOLO model has three main components. The backbone, neck, and the prediction head. Each of these components will be discussed later in detail. Once the areas of interest (potential breast masses) have been identified, these regions are cropped from the original images, preparing them for the next stage.

Following the object detection phase, the regions of interest, which are the cropped breast mass regions, are directed into the classification stage, as it can be seen in Figure 5. The objective of the classification stage is to distinguish between benign and malignant breast masses. In order to do this, we used transfer learning, specifically through the fine-tuning of pre-trained models. We used a range of pre-trained models for the classification task including DenseNet 121, Inception V3, VGG 16, AlexNet, ResNet 18, and ResNet 50.

3.4. YOLO

YOLO (You Only Look Once) object detection algorithm is a widely used one-stage object detection algorithm. As opposed to two-stage detectors, YOLO performs object detection in one go. The input image is split into grids, and each cell in the grid is responsible for predicting objects within it. It simultaneously predicts the bounding boxes and class probabilities for these boxes. (Redmon et al., 2016)

Two-stage detectors, such as R-CNN, Fast R-CNN, and Faster R-CNN, approach object detection in two primary steps. The first stage involves generating a set of proposal regions within the image where the object might be located, and this is usually done using a region proposal network. Once these candidate regions are proposed, the second stage involves extracting features from them and classifying them to identify the object. (Du et al., 2020)

3.5. YOLOv5

Yolov5, as its former variant architecture, uses the darknet architecture. Mainly the model's architecture is a composition of the backbone, neck, and head. The backbone is a CSP-Darknet53 that is responsible for extracting relevant features from the input image. Then the extracted features will pass to the next component of the architecture. The neck, SPPF (Spatial Pyramid Pooling Fast) serves as the transitional component. It uses multiple convolutional layers and pooling layers to create a multiscale feature map. This is useful for detecting objects of different sizes in the image. The head, using the previous YOLOV3 head, is the prediction layer where the actual object detection occurs. It uses the multiscale feature map generated by the neck to make predictions about the presence and location of objects in the input image. (Jocher et al., 2020) At this stage bounding box coordinates for each detected object and the object class is predicted. It also offers

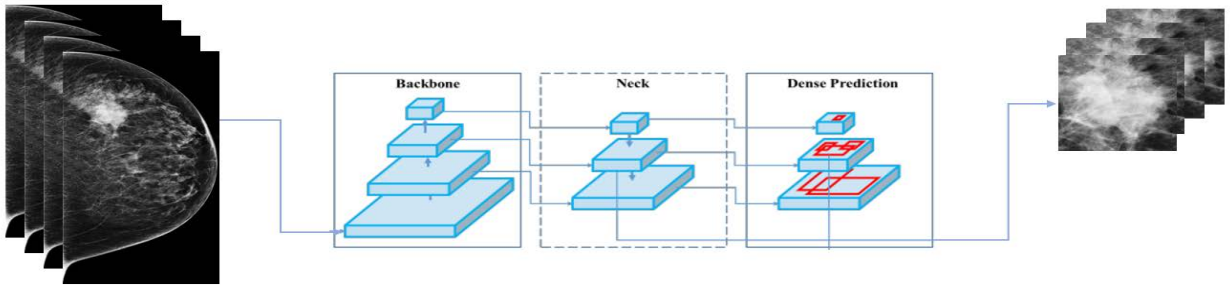


Figure 4: Proposed method for breast mass detection

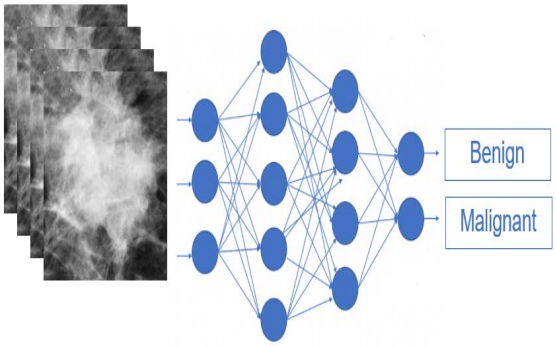


Figure 5: Proposed method for classification task

multiple versions of the model YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x that vary in size and computational requirements.

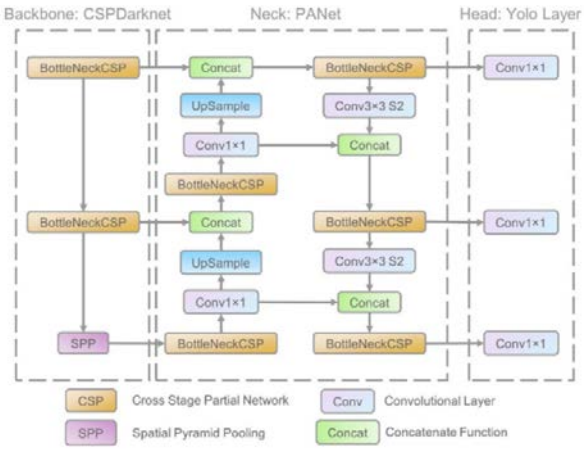


Figure 6: YOLOV5 Architecture

In this study, we used YOLOv5 models and, specifically, yolov5s, yolov5m, yolov5l and yolov5m6 pretrained models, along with RadImageNet weights. Each of these models is differentiated by their computational size and capacity, with 's' referring to a small model, 'm' to medium, 'l' to large, and 'm6' being a medium-sized model in the updated version six.

As an extension to the foundational YOLOv5 model, we also used a variant incorporating Transformer architectures right on top of the conventional YOLOv5 structure. YOLOv5 Transformer Prediction Head (TPH) design modification helps enhance the overall performance of the object detection model by allowing it to more effectively learn and process contextual information in the image data. Due to their attention mechanism, transformers can capture interactions between distant pixels, thereby improving the model's ability to detect objects, especially in complex scenes where traditional convolutional methods may struggle. (Zhu et al., 2021)

Yolov5 Models	Parameters	
YOLOV5S	Image size	1024
	Batch size	16
	Epochs	100
YOLOV5M	Image size	1024
	Batch size	16
	Epochs	100
YOLOV5M6	Image size	640
	Batch size	32
	Epochs	100
	Freeze	12 layers
YOLOV5M6	Image size	1280
	Batch size	8
	Epochs	100
YOLOV5M6	Image size	1536
	Batch size	2
	Epochs	100
	Configuration	TPH
YOLOV5L	Image size	1024
	Batch size	16
	Epochs	100

Alongside various YOLOv5 models, we used RadImageNet weights. The YOLOv5 models are typically trained on the MS COCO dataset, which comprises a wide variety of general-purpose images. Whereas RadImageNet is exclusively trained on medical image datasets. This characteristic led us to assume that using RadImageNet could potentially enhance our model's performance in detecting breast masses, as it has already been exposed to and trained on medical data during its

development phase. By leveraging these weights, the model is provided with a more suitable initial configuration, setting it on the right path toward learning the features of breast masses. This pre-training could reduce the necessary training time and potentially improve the final performance of the model.

Our study further tested our hypothesis about the benefit of using RadImageNet weights with more specific models: ResNet50, DenseNet121, and InceptionV3. These models are trained on a small subset of around 1.4 million medical images (Mei et al., 2020). We used these models' weights instead of training them from scratch or using the pre-trained weights from MS COCO dataset.

RadImageNet	Yolov5 Models	Parameters
InceptionV3	YOLOV5M	Image size- 640 Batch size - 16 Epochs- 300
DenseNet121	YOLOV5M	Image size- 640 Batch size - 16 Epochs- 300
ResNet50	YOLOV5M	Image size- 640 Batch size - 16 Epochs- 300

3.6. YOLOV5-Transfomer Prediction Head

The Transformer Prediction Head variant of YOLOv5 is another variation and a modified version of the YOLOV5 models' heads. In a nutshell, it integrates a transformer prediction head in place of the conventional convolutional layers in YOLOv5's prediction stage. This was inspired by the recent successes of transformers compared to CNN. The Transformer Prediction Head follows the same core architecture as YOLOv5, retaining the basic features of the YOLO family. It comprises of three main components: a backbone for feature extraction, a neck for multi-scale feature aggregation, and a prediction head for object detection. Like other YOLO models, the backbone is responsible for extracting features from the input images. It is generally a deep convolutional neural network (CNN) that can process an image and output a set of feature maps that encapsulate the salient information in the image. Followed by the neck, where it merges the feature maps from the backbone at multiple scales. This enables the model to detect objects of various sizes present in the image. The prediction head is where this variant diverges from the conventional YOLO architecture. Instead of using convolutional layers, it uses a transformer network for prediction. The transformer prediction head uses self-attention mechanisms to model the various relationships between different image parts. This helps understand complex spatial dependencies and improves the model's ability to locate and classify objects in the image (Zhao et al., 2023).

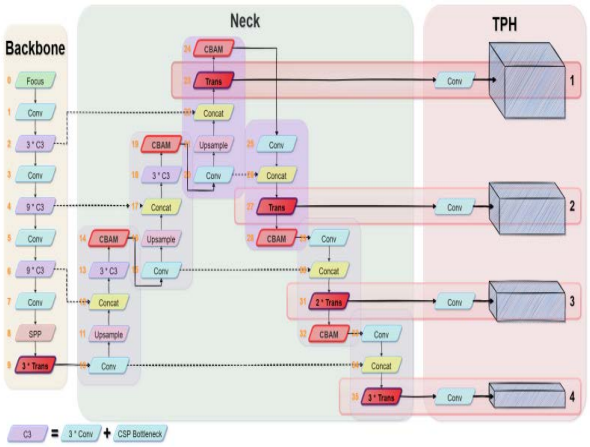


Figure 7: YOLOV5 with four transformer prediction head architecture

One of the significant advantages of using a transformer in the prediction head is its ability to model long-range dependencies. Convolutional layers typically focus on local features, and while pooling layers can help capture larger contexts, they may miss out on specific long-range relationships. On the other hand, transformers are explicitly designed to handle these kinds of dependencies, making them a powerful tool for tasks like object detection.

3.7. YOLOV6

YOLOv6 is another addition to the YOLO series. It aims to optimize the YOLO framework further by incorporating different scales among various models. Smaller models use a single-path backbone, while larger models were built upon efficient multi-branch blocks. This strategy is aimed at optimizing the trade-off between speed and accuracy. Another characteristic of YOLOv6 is the use of a self-distillation strategy (Li et al., 2022). The strategy was utilized for both the classification task and the regression task. The aim was to enable the student model to learn more efficiently from the teacher and labels during all training phases.

In terms of network design, the backbone of YOLOv6 differed based on the size of the models. For smaller models, the backbone during the training phase was the RepBlock. Larger models use a CSP block called CSP-StackRep Block. The neck is built on the Path Aggregation Networks(PAN) model. Followed by Efficient Decoupled Head.

3.8. YOLOV7

The YOLOv7 series focuses on optimizing both the architecture and the training process to maximize performance and efficiency. It uses Extended Efficient

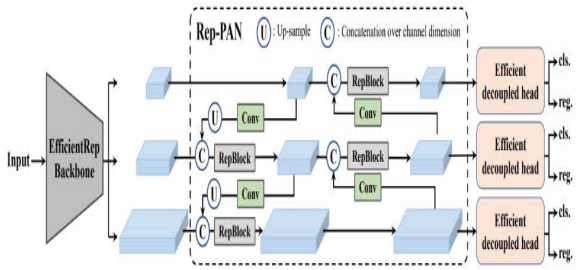


Figure 8: YOLOV6 Architecture

Layer Aggregation Networks (E-ELAN). E-ELAN is designed to maximize learning capabilities without disrupting the gradient path, which can improve efficiency and performance. This architecture employs an expand, shuffle, merge approach to guide the network's learning (Wang et al., 2023). Regarding model scaling, YOLOv7 introduces a new method specifically designed for concatenation-based models. This approach is designed to balance the impacts of scaling on different aspects of the model, maintaining optimal structure and performance across different scales. It also implements a unique deep supervision strategy with two kinds of heads: an auxiliary head to assist in training and a lead head responsible for the final output. This arrangement helps improve the model's overall performance.

3.9. YOLOV8

YOLOv8 is another new state-of-the-art object detection model. Its architecture is divided into two primary components: the backbone and the head. The backbone of YOLOv8 is an adaptation of the CSPDarknet53 architecture. It uses a network of 53 convolutional layers in cross-stage partial connections. In the head, the model integrates a self-attention mechanism. It evaluates the relative importance of various features, adjusting its attention according to the relevance of these features to the task. This selective attention allows for more refined object detection, as the model can better detect objects of interest from background noise. YOLOv8 also enhances spatial attention, feature fusion, and context aggregation modules. Spatial attention helps the model focus on specific locations in the image space likely to contain objects of interest. Feature fusion enables the model to combine information from different types and levels of features, and context aggregation allows the model to integrate contextual information from the surrounding image, improving its ability to differentiate objects from their backgrounds.

3.10. Evaluation Metrics

There are various performance evaluation metrics for object detection tasks, Several evaluation metrics are used to give a comprehensive understanding of the model's performance (Padilla et al., 2021). Our analysis for the breast mass detection methods is based on some of these commonly known metrics applied to test datasets. We mainly used precision, recall, and mean average precision (mAP) along with the measure of true positive rate(TPR) at a threshold of 0.85 false positives per image (FPPI)

Below, we will describe four commonly used evaluation metrics in object detection tasks: Precision, Recall, Mean Average Precision (mAP), and Intersection Over Union (IoU).

Precision measures the accuracy of the detected instances. It quantifies the number of correct positive predictions made. Specifically, it is the ratio of the correctly predicted positive observations (True Positives) to the total predicted positive observations, which includes both True Positives and False Positives (incorrectly identified as positive).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall/Sensitivity/True Positive Rate quantifies the model's ability to find all the relevant instances in a dataset. It is the ratio of correctly predicted positive observations to all actual positive observations in the data.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Mean Average Precision (mAP) is a single number summary of the average precision at varying recall levels. It's one of the most important evaluation metrics for checking the accuracy of an object detection model. mAP considers both Precision and Recall to compute the score. For every predicted bounding box, the Average Precision (AP) is calculated, and mAP is the mean of APs for all classes.

Intersection Over Union (IoU) is a measure of the overlap between ground truth (B_{pt}) and predicted bounding (B_{gt}) boxes. It is used as a criterion to decide if a prediction is a True Positive, False Positive, or False Negative. The IoU of two bounding boxes is calculated as the area of their overlap divided by the area of their union. (Padilla et al., 2021)

$$\text{IOU} = \frac{\text{area}\left(B_p \cap B_{gt}\right)}{\text{area}\left(B_p \cup B_{gt}\right)}$$

4. Results

In this section, we will discuss in detail the performance of the different YOLO models and versions that we have utilized in the process of detecting breast masses. we will deal with the specific performance of

each of these models in our task and analyze the detection accuracy, classification precision, recall, and F1-score. The performance of these models will be compared with each other, providing us with a clear idea about which model performs the best in the context of breast mass detection and classification.

4.1. Breast Mass Detection

The results obtained from the various trained models of YOLOV5 for breast mass detection are summarized in the table below. The four performance metrics used in the analysis include Precision, Recall, mAP@50 (mean average precision at Intersection over Union (IoU) over 0.50), and TPR at FPPI of 0.85 (True Positive Rate at a False Positive Per Image rate of 0.85). In table 1, YOLOV5m performed best in terms of precision, achieving a score of 0.694. Precision measures the proportion of correctly predicted positive observations to the total predicted positives. Regarding recall metrics, the DenseNet121 model trained on the RadImageNet dataset with the YOLOV5m architecture performed the best, achieving a recall of 0.713. Recall measures the proportion of correctly identified positive cases from all actual positive cases. Looking at mAP@50, the DenseNet121 model trained on the RadImageNet dataset with YOLOV5m again stands out, achieving a mAP@50 of 0.718. The true positive rate at an FPPI of 0.85 is highest for the DenseNet121 with a value of 0.97. In conclusion, based on the metrics, the DenseNet121 model trained on the RadImageNet dataset with YOLOV5m seems to outperform the other models in this task. This model appears to provide a good balance of precision and recall, leading to better overall performance in detecting and classifying breast masses.

Our experiments with the DenseNet121 model trained on the RadImageNet dataset and subsequently applied to breast mass detection have shown good results. A visual representation of these results is illustrated in the Figure 10.

In our experiment, we examined the performance of the YOLOv5 model with a Transformer prediction head, coupled with RadImageNet DenseNet 121 weights, to evaluate its efficiency in detecting breast masses, as shown in Figure 11. We achieved a True Positive Rate (TPR) of 0.89 at a False Positive Per Image (FPPI) threshold of 0.85.

Yolov5 transformer prediction head model result, when compared with another variant of the YOLOv5 model, which utilizes the DenseNet121 weights from RadImageNet, shows a notable performance difference, as can be seen in Figure 12. The YOLOv5 model, when paired with DenseNet121 weights, outperformed the Transformer prediction head model variant by yielding the highest performance in our tests. It achieved a TPR of 0.97 at the same FPPI threshold of 0.85.

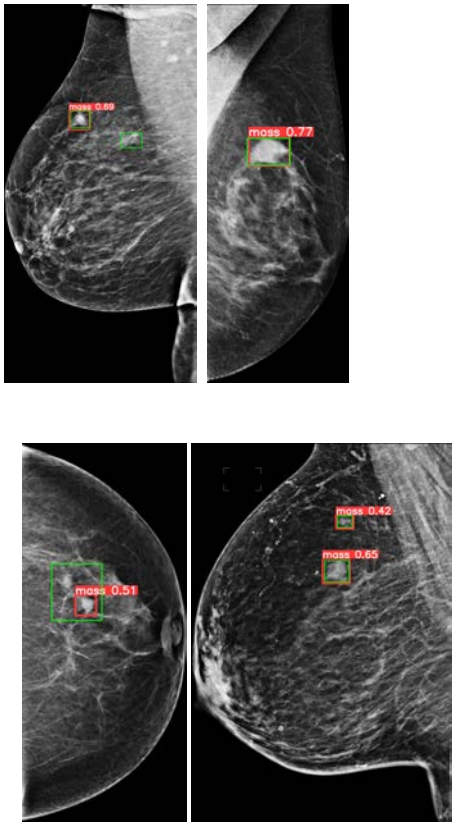


Figure 9: Example of detected breast masses with yolov5m with RadImageNet-DenseNet121 weights

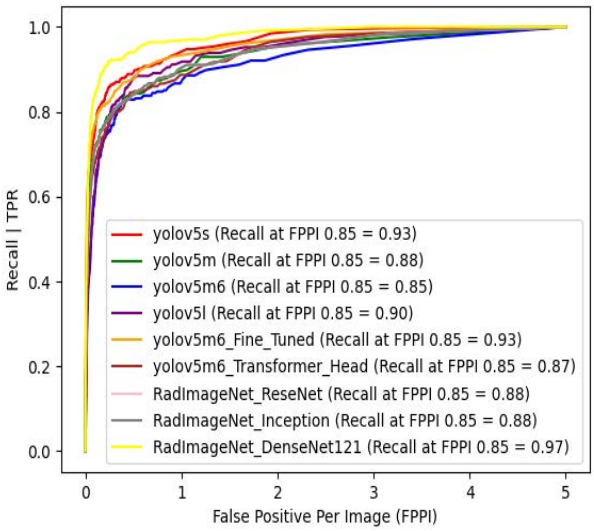


Figure 10: FROC Curve for YOLOV5 all trial

	Precision	Recall	mAP@50	TPR at FPPI 0.85
YOLOV5s	0.619	0.537	0.556	0.93
YOLOV5m	**0.694**	0.609	0.600	0.88
YOLOV5l	0.616	0.544	0.533	0.90
YOLOV5m6	0.636	0.606	0.588	0.85
YOLOV5m6 Frozen Layer	0.641	0.529	0.534	0.93
YOLOV5m6 Transformer Head	0.652	0.551	0.568	0.87
RadImageNet InceptionV3 YOLOV5m	0.605	0.606	0.557	0.88
RadImageNet ResNet50 YOLOV5m	0.605	0.606	0.557	0.88
RadImageNet DenseNet121 YOLOV5m	0.678	**0.713**	**0.718**	**0.97**

Table 1: YOLOV5 all trained models result

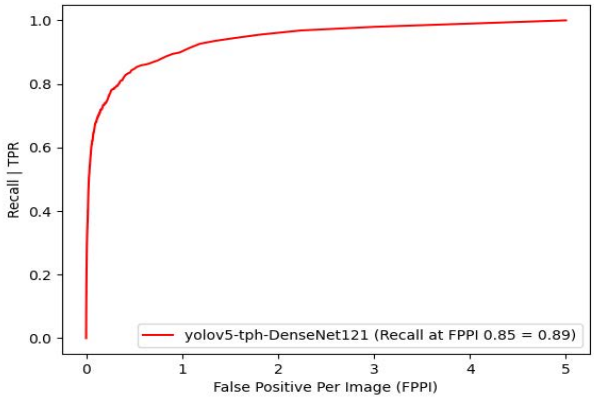


Figure 11: FROC Curve for YOLOV5-transformer prediction head trial

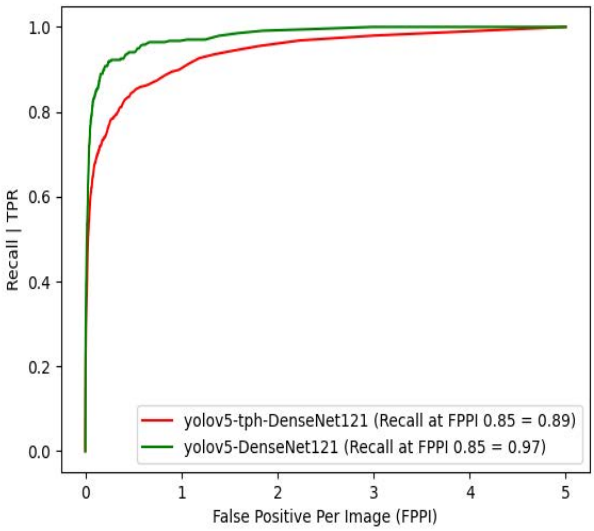


Figure 12: Comparision between Yolov5 with RadImageNet-DenseNet121 weights vs Yolov5 transformer prediction head along with RadImageNet-DenseNet121 weights

This performance gap highlights model architecture's significant impact on detection performance. It indicates that the YOLOv5 model combined with DenseNet121 weights performs more efficiently in detecting breast masses, providing more accurate results than the variant that uses a transformer prediction head. It's important to note that while transformers have shown promising results, their performance in object detection tasks might vary depending on the dataset and problem context.

The area under the Precision-Recall curve mAP of the yolov5 transformer prediction head along with RadImageNet-DenseNet121 weight yields 0.51 as it can be shown in Figure 10

In our experiment with the YOLOv6 object detection model, we used the YOLOv6l6 model, which is better performing due to its larger architecture. The original YOLOv6l6 had reached a mean Average Precision (mAP) of 57.2 at an Intersection over Union (IoU) of 0.5 on a COCO dataset. After fine tuning the model and running it on for 100 epochs we got an mAP of 0.625 at 0.5 IoU.

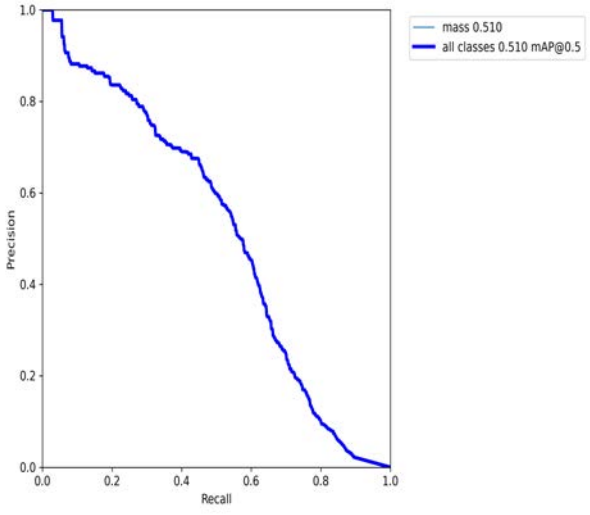


Figure 13: Precision vs Recall curve for Yolov5 transformer prediction head on test dataset

	Precision	Recall	mAP@50	TPR at FPPI 0.85
YOLOV5-Transformer Prediction Head	0.558	0.540	0.510	0.89

Table 2: YOLOV5-Transformer Prediction Head result on test dataset

	Recall	mAP@50
YOLOV6l6	0.601	0.625

There are six models provided by YOLOv7, all of which were trained on the MS COCO dataset. Out of these, we choose the YOLOv7-E6E model. It's the largest as well as it is also more accurate compared to the rest of the models, which is significantly important when it comes to choosing a model for this study. It is important to highlight that the performance benefits of YOLOv7-E6E come at the expense of speed. As our work with the model demonstrated, it is relatively slower than the other options available. For instance, a run of 100 epochs took over 72 hours to complete. However, given the superior accuracy of the YOLOv7-E6E, the trade-off between speed and accuracy was acceptable in the context of this study.

We also trained the YOLOv7-E6E model with an initialized weight from RadImageNet - DenseNet121. This strategy was adopted to evaluate and compare the performance of the YOLOv7-E6E model against the RadImageNet dataset. The results show that the performances of the RadImageNet - DenseNet121 and the YOLOv7-E6E models were closely matched. However, the YOLOv7-E6E model exhibited slightly better performance.

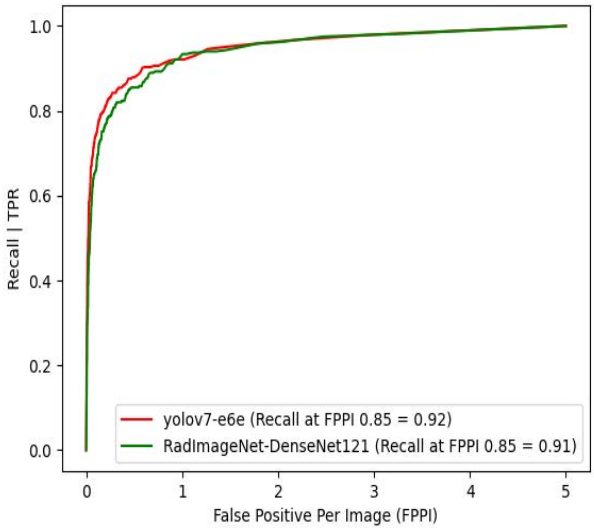


Figure 14: FROC Curve for YOLOV7 trial

The Precision-Recall curve is a vital tool for understanding the performance of an object detection model. It is a plot of the precision and the recall for different thresholds. A model with perfect precision and recall would achieve a point at the top right corner of the plot, indicating that it has perfectly classified all positive instances without making any false-positive errors. In practice, however, models usually exhibit a trade-off between precision and recall.

The area under the Precision-Recall curve (AUPRC), also known as mean average precision (mAP), is a single-value metric that summarizes the model's overall quality across all thresholds. It considers both precision and recall at every possible threshold and effectively summarizes the balance between them.

In the context of YOLOv7-E6E's performance, as seen in Figure 15, achieving an mAP of 0.618 shows that the model has a good balance of precision and recall and is relatively reliable in identifying breast masses. However, there is still room for improvement as the mAP value is not very close to 1.0.

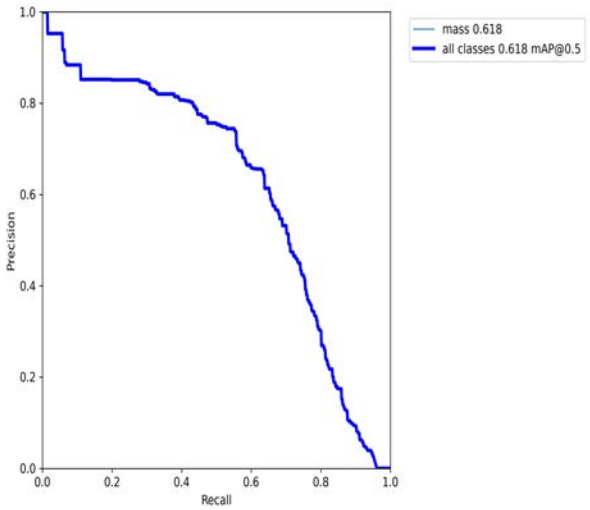


Figure 15: Precision vs Recall curve for Yolov7-e6e on test dataset

After integrating the DenseNet121 weights trained on the RadImageNet dataset with the YOLOv7-E6E model, a slight decrease in the mean average precision (mAP) value was observed during our experiments, as shown in Figure 16.

For the yolov8 trial we used a yolov8X and yolov8l models, yolov8x is an extra-large size, although it leads to a relatively slower computational speed, often provides higher accuracy. The YOLOv8X model's performance has been evaluated, and it has achieved a 0.53 mAP score at 0.5 IoU on the COCO validation dataset.

	Precision	Recall	mAP@50	TPR at FPPI 0.85
YOLOV7-e6e	0.655	**0.632**	**0.618**	**0.92**
RadImageNet DenseNet121 YOLOV7-e6e	**0.677**	0.585	0.600	0.91

Table 3: YOLOV7 trained models result on test dataset

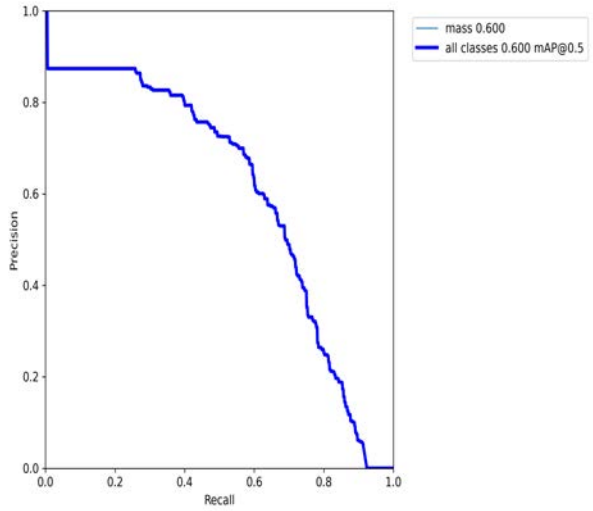


Figure 16: Precision vs Recall curve for Yolov7 with DenseNet121 weights

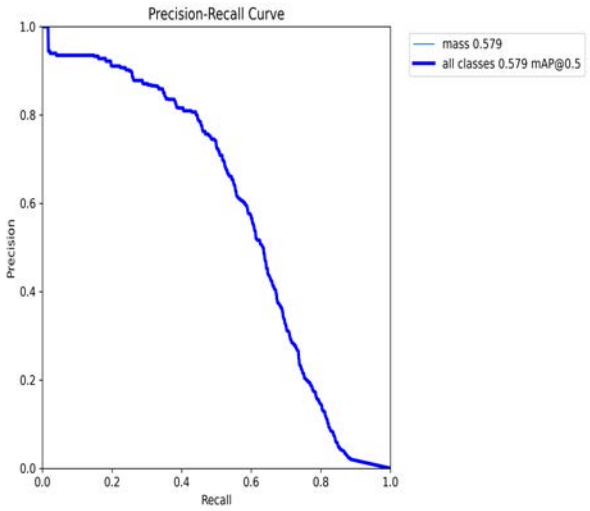


Figure 17: Precision vs Recall curve for Yolov8x model

When we applied this model to our dataset, the mAP results at 0.5 IoU reached 0.579. Figure 17 visually represents the Precision vs Recall curve for the YOLOv8X model on our dataset.

Additionally, the yolov8l, a large-sized model, performance is in close comparison to the yolov8x model with an mAP of 0.563 at 0.5 IoU. The precision vs recall curve shows the results in Figure 18.

4.2. Breast Mass Classification

This study follows a two-step pipeline, starting with detecting any masses present in the breast, then the classification of these detected masses into benign or malignant categories. This approach allows for a more efficient and effective diagnosis, facilitating early detection and treatment of potential breast cancer cases.

In our approach, we utilized the YOLOV5m model, trained with RadImageNet DenseNet121 weights, to initially detect breast masses. This particular model was chosen due to its outstanding performance compared to other models. As demonstrated in our detection analysis, this model exhibited high precision in detecting breast masses, making it an ideal candidate for this critical first step.

The identified breast masses were extracted from the original mammogram images after the detection phase. This process involves cropping the image around the region identified as a mass by the YOLOV5m detector. These cropped sections, each containing a single mass,

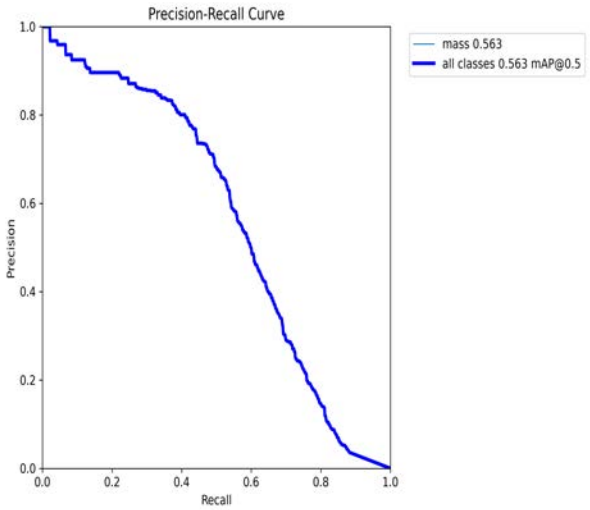


Figure 18: Precision vs Recall curve for Yolov8l model

	Precision	Recall	mAP@50
YOLOV8X	0.690	0.523	0.579
YOLOV8L	0.673	0.506	0.563

Table 4: YOLOV8X result in test dataset

were then prepared for the next process phase: classification.

With a collection of cropped images of detected breast masses, we moved into the classification phase. The purpose of this step is to classify each detected mass as either benign (non-cancerous) or malignant (cancerous). This information is crucial to the subsequent medical procedures, guiding clinicians in choosing treatment and intervention strategies.

The classification model has been trained to recognize the distinguishing features between benign and malignant breast masses, thereby accurately classifying new instances. By feeding the cropped images from the detection phase into this classifier, we were able to generate a robust, two-step diagnostic tool that both identifies and categorizes breast masses.

In our study, we faced a substantial class imbalance problem that could potentially impact the performance of our classification model. This issue is largely attributed to the nature of our dataset, which comprises a significant number of benign breast masses compared to malignant cases. Out of the 3363 detected breast masses, only 312 were benign, while the rest 3051 were malignant cases. Such class imbalance can lead to a bias in the classifier towards the majority class, in this case, the malignant masses.

The class imbalance problem is a well-known challenge in the field of deep learning, especially in medical imaging, where the number of positive cases can be considerably lower than the number of negative cases. This imbalance can introduce a bias towards the majority class during the training phase, leading to a model that performs poorly on the minority class. This is a significant concern, as the misclassification of malignant masses could have severe consequences in a clinical setting.

Addressing this problem required implementing several strategies to ensure our model performed optimally despite the imbalance. One of the approaches we employed was applying class weights during the training phase. Class weights are a powerful tool in machine learning that can help to balance out the influence of each class during training. By assigning higher weights to the minority class (benign cases in our dataset), we can increase their impact on the model's learning process, thereby reducing the bias towards the majority class. Another strategy we implemented was using a regularization method known as weight decay. Regularization is a technique used to prevent overfitting, which is a common problem when a model learns to perform too well on the training data, and as a result, performs poorly on unseen data. Weight decay works by adding a penalty to the loss function based on the magnitude of the weights in the model. This encourages the model to learn simpler decision boundaries, leading to a model that generalizes better to new data.

The combined implementation of class weights and weight decay proved to be a promising solution to our class imbalance problem. Applying these methods resulted in a less biased model towards the majority class and performed better on the minority class. This approach, while not entirely eliminating the imbalance, significantly resolved its impact on our model's performance.

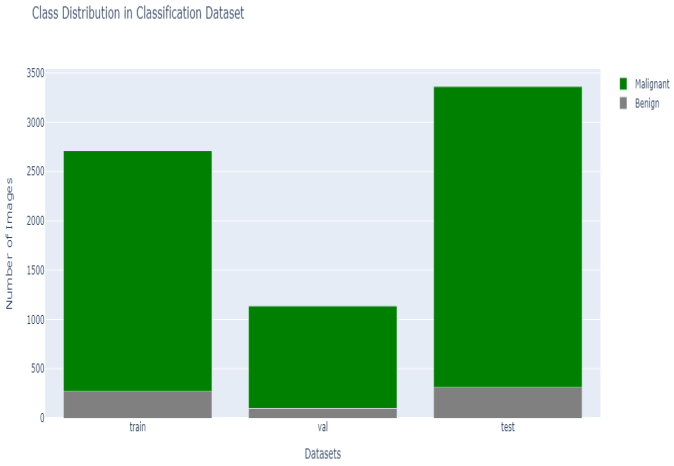


Figure 19: Breast mass classification dataset

As can be seen in Figure 19, there is a significant class imbalance in our mass classification dataset. We based our classification task on a dataset derived from the ground truth since there were instances of undetected masses left behind from the detection stage. Our original dataset comprised 3849 breast masses, out of which we were successful in detecting only 3363 masses. Consequently, due to the incomplete detection, we decided to use the entire dataset of 3849 masses for our training and validation. The reason behind this division was to create a robust model that could generalize well in the provided data. This division resulted in a training dataset comprising 2711 images and a validation set with 1138 images. For the

testing phase of our study, we used the data derived from the previous detection pipeline, totaling 3363 images. Therefore, the testing data represents the actual conditions in which the model would work.

Class Distribution	
Train	Benign - 271 Malignant - 2440
Validation	Benign - 101 Malignant - 1037
Test	Benign - 312 Malignant - 3051

5. Discussion

In this study, we carried out a two-step approach for breast mass detection and classification. Primarily, we utilized a variety of YOLO-based architecture variants for our detection task. By effectively integrating both pre-trained weights on the MS COCO dataset and leveraging the RadImageNet dataset weights across several CNN architectures, including InceptionV3, DenseNet121, and ResNet50. For the selection of the detection models, we evaluated an array of YOLO variants, including YOLOV5, YOLOV6, YOLOV7, and YOLOV8. Among the evaluated models, DenseNet121, initialized with RadImageNet weights and integrated with the YOLOV5m architecture, stood out in terms of performance. In this experiment, we got 0.97 TPR at 0.85 FPPI. This is a satisfactory result balancing between detection accuracy and computation efficiency, making it an ideal choice for subsequent classification tasks.

The subsequent breast mass classification phase was predicated on the detected regions by the best-performing detection model. Each of the detected breast masses was further cropped and fed into the classifier model as a test case. We leveraged transfer learning architectures to train the classification task. As there were some missing masses left undetected from the detection model. We opted to use the dataset with the annotated ground truth value for training and validation.

Furthermore, the comprehensive evaluation metrics deployed in this study support the ability to assess model performance. Precision, Recall, Mean Average Precision (mAP), True Positive Rate at False Positive Per Image threshold of 0.85, and Intersection over Union (IoU) were calculated to offer a holistic view of the models' effectiveness. These metrics provided a robust framework for model evaluation, ensuring that the assessment was comprehensive, fair, and objective.

The classification stage of our study posed a challenge due to a significant class imbalance between the benign and malignant classes. To resolve this issue, we implemented class weighting during the training process. This approach assigns different weights to the classes inversely proportional to their frequency. As a result, the model pays more attention to the less-represented class during the learning process.

Alongside class weighting, we also applied weight decay regularization, a common method to prevent overfitting in deep learning models. This technique adds a penalty to the loss function based on the size of the weights, discouraging the model from learning overly complex patterns that may not generalize well to unseen data. Despite these adjustments, the class imbalance issue wasn't fully resolved. Yet, we were able to achieve a reasonable performance with the VGG16 model, which attained an accuracy of 0.924 a comparative result in other evaluation metrics.

6. Conclusions

In conclusion, our work showcases a two-stage approach for detecting and classifying breast masses, demonstrating the potential for integrating object detection models, such as YOLO and it's variants for breast mass detection. Moreover, our work underlines the importance of using models trained specifically on medical images. In our study, the use of the RadImageNet model, which is specially designed and trained on radiological images, stands out as a particularly effective strategy. The positive impact of using such domain-specific models has significant implications for future medical imaging studies.

For our breast mass detection task, we used transfer learning. The models were fine-tuned for our dataset, allowing us to adapt these high-performing models into the specifics of our task. Although we faced a high class imbalance in our breast mass detection dataset, where the benign cases were significantly outnumbered by malignant cases. we mitigate the problem using class weighting and regularization approaches.

Acknowledgments

I would like to extend my heartfelt thanks to my supervisor, Dr.Mario Molinara. His valuable guidance, unwavering support, and constructive suggestions played a vital role in shaping this research. His expertise and insights have been invaluable to me. I'm grateful to the Università degli Studi di Cassino for providing me the opportunity to pursue my master's thesis. I would alsolike to thank Laura Morone for her assistance during my onboarding process.

	Accuracy	F1-Score	Precision	Recall
DenseNet121	0.896	0.602	0.655	0.582
InceptionV3	0.893	0.634	0.661	0.617
AlexNet	0.901	0.556	0.647	0.544
VGG-16	0.924	0.731	0.793	0.695
ReseNet18	0.881	0.620	0.631	0.611
ResNet 50	0.888	0.606	0.634	0.591
EfficientNetB0	0.906	0.505	0.672	0.513

Table 5: Transfer Learning for classification task

I am deeply grateful to the Medical Imaging and Application (MAIA) program and all the partner universities. Their collective effort and dedication have created an engaging and challenging environment where I could thrive. Last, but not the least, I want to thank my family and friends for their support and encouragement.

References

Agarwal, R., Díaz, O., Yap, M.H., Lladó, X., Martí, R., 2020. Deep learning for mass detection in full field digital mammograms. Computers in Biology and Medicine 121, 103774. doi:10.1016/j.compbiomed.2020.103774.

Akselrod-Ballin, A., Chorev, M., Shoshan, Y., Spiro, A., Hazan, A., Melamed, R., Barkan, E., Herzel, E., Naor, S., Karavani, E., et al., 2019. Predicting breast cancer by applying deep learning to linked health records and mammograms. Radiology 292, 331–342. doi:10.1148/radiol.2019182622.

Al-Masni, M.A., Al-Antari, M.A., Park, J.M., Gi, G., Kim, T.Y., Rivera, P., Valarezo, E., Choi, M.T., Han, S.M., Kim, T.S., 2018. Simultaneous detection and classification of breast masses in digital mammograms via a deep learning yolo-based cad system. Computer methods and programs in biomedicine 157, 85–94.

Betancourt Tarifa, A.S., Marrocco, C., Molinara, M., Tortorella, F., Bria, A., 2023. Transformer-based mass detection in digital mammograms. Journal of Ambient Intelligence and Humanized Computing 14, 2723–2737.

Du, L., Zhang, R., Wang, X., 2020. Overview of two-stage object detection algorithms, in: Journal of Physics: Conference Series, IOP Publishing. p. 012033.

Evans, K.K., Haygood, T.M., Cooper, J., Culpan, A.M., Wolfe, J.M., 2016. A half-second glimpse often lets radiologists identify breast cancer cases even when viewing the mammogram of the opposite breast. Proceedings of the National Academy of Sciences 113, 10292–10297.

Halling-Brown, M.D., Warren, L.M., Ward, D., Lewis, E., Mackenzie, A., Wallis, M.G., Wilkinson, L.S., Given-Wilson, R.M., McAvinchey, R., Young, K.C., 2021. Optimam mammography image database: A large-scale resource of mammography images and clinical data. Radiology: Artificial Intelligence 3. doi:10.1148/ryai.2020200103.

Jocher, G., Changyu, L., Hogan, A., Yu, L., Rai, P., Sullivan, T., et al., 2020. ultralytics/yolov5: Initial release. Zenodo .

Lévy, D., Jain, A., 2016. Breast mass classification from mammograms using deep convolutional neural networks. arXiv preprint arXiv:1612.00542 .

Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Ke, Z., Li, Q., Cheng, M., Nie, W., et al., 2022. Yolov6: A single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976 .

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer. pp. 740–755.

Mattiuzzi, C., Lippi, G., 2019. Current cancer epidemiology. Journal of epidemiology and global health 9, 217.

Mei, X., Lee, H.C., Diao, K.y., Huang, M., Lin, B., Liu, C., Xie, Z., Ma, Y., Robson, P.M., Chung, M., et al., 2020. Artificial intelligence–enabled rapid diagnosis of patients with covid-19. Nature medicine 26, 1224–1228.

Mei, X., Liu, Z., Robson, P.M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K.E., Yang, T., et al., 2022. Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. Radiology: Artificial Intelligence 4, e210315.

Padilla, R., Passos, W.L., Dias, T.L.B., Netto, S.L., da Silva, E.A.B., 2021. A comparative analysis of object detection metrics with a companion open-source toolkit. Electronics 10. URL: https://www.mdpi.com/2079-9292/10/3/279, doi:10.3390/electronics10030279.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788.

Rodriguez-Ruiz, A., Lång, K., Gubern-Merida, A., Broeders, M., Gennaro, G., Clauser, P., Helbich, T.H., Chevalier, M., Tan, T., Mertelmeier, T., et al., 2019. Stand-alone artificial intelligence for breast cancer detection in mammography: Comparison with 101 radiologists. JNCI: Journal of the National Cancer Institute 111, 916–922. doi:10.1093/jnci/djy222.

Ryspayeva, M., Molinara, M., 2022. Breast Mass Detection and Classification Using Transfer Learning. Master's thesis. Universit'a degli studi di Cassino e del Lazio Meridionale.

Sampat, M.P., Markey, M.K., Bovik, A.C., et al., 2005. Computer-aided detection and diagnosis in mammography. Handbook of image and video processing 2, 1195–1217.

Sechopoulos, I., Teuwen, J., Mann, R., 2021. Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art, in: Seminars in Cancer Biology, Elsevier. pp. 214–225.

Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M., 2016. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. IEEE Transactions on Medical Imaging 35, 1285–1298. doi:10.1109/tmi.2016.2528162.

Siegel, R.L., Miller, K.D., Wagle, N.S., Jemal, A., 2023. Cancer statistics, 2023. CA: A Cancer Journal for Clinicians 73, 17–48. doi:10.3322/caac.21763.

Tabár, L., Dean, P.B., Chen, T.H., Yen, A.M., Chen, S.L., Fann, J.C., Chiu, S.Y., Ku, M.M., Wu, W.Y., Hsu, C., et al., 2018. The incidence of fatal breast cancer measures the increased effectiveness of therapy in women participating in mammography screening. Cancer 125, 515–523. doi:10.1002/cncr.31840.

Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M., 2023. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in: Proceedings of the IEEE/CVF Conference on Computer

Vision and Pattern Recognition, pp. 7464–7475.

Yan, Y., Conze, P.H., Lamard, M., Quellec, G., Cochener, B., Coatrieux, G., 2021. Towards improved breast mass detection using dual-view mammogram matching. Medical Image Analysis 71, 102083. doi:10.1016/j.media.2021.102083.

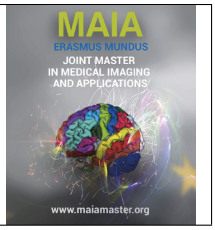
Zahoor, S., Lali, I.U., Khan, M.A., Javed, K., Mehmood, W., 2020. Breast cancer detection and classification using traditional computer vision techniques: a comprehensive review. Current medical imaging 16, 1187–1200.

Zhao, Q., Liu, B., Lyu, S., Wang, C., Zhang, H., 2023. Tph-yolov5++: Boosting object detection on drone-captured scenarios with cross-layer asymmetric transformer. Remote Sensing 15. URL: https://www.mdpi.com/2072-4292/15/6/1687, doi:10.3390/rs15061687.

Zhu, X., Lyu, S., Wang, X., Zhao, Q., 2021. Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios. CoRR abs/2108.11539. URL: https://arxiv.org/abs/2108.11539, arXiv:2108.11539.

Medical Imaging and Applications

Master Thesis, June 2023



Low-dose CT Reconstruction with Active Learning and Implicit Neural Representation

Manasi Kattel, Francisco Vasconcelos, Binod Bhattarai

Wellcome/EPSRC Centre for Interventional and Surgical Sciences(WEISS), University College London, UK

Abstract

Computed tomography (CT) is a vital medical imaging modality that provides detailed cross-sectional images of the human body. CT plays a crucial role in clinical diagnosis, treatment planning, and disease monitoring. However, CT imaging involves exposing patients to ionizing radiation, which raises concerns regarding potential health risks associated with repeated exposure. To address these concerns, we focus on the significance of low-dose CT reconstruction, which aims to minimize radiation dose while maintaining high-quality image reconstruction. Our proposed framework utilizes an Implicit Neural Representation (INR) approach combined with active projection sampling techniques to improve the accuracy and efficiency of low-dose and highly ill-posed CT reconstruction. By combining Fourier feature encoding and incorporating prior terms into a Multi-Layer Perceptron (MLP) fitting process, we achieve optimal results in scenarios with limited views. Consequently, we select INR as the foundation for our reconstruction approach and proceed to investigate active sampling methods for sampling projections. Our active sampling study focuses on comparing greedy approaches for projection sampling and highlights the advantages of non-uniform sampling over uniform sampling. Through extensive evaluations, we assess the performance of our models on both a Shepp-Logan type phantom dataset and a Low-Dose Parallel Beam (LoDoPaB)-CT dataset, specifically targeting sparse view cases as low as 8 views for the Phantom dataset and 25 views for the LoDoPaB dataset. Furthermore, we reveal that the angles chosen for optimal reconstructions exhibit discernible patterns, suggesting a link with the underlying anatomical structures being reconstructed.

Keywords: Inverse problems, Low dose CT Reconstruction, Implicit Neural Representations, Active Learning

1. Introduction

Computed tomography (CT) is a non-invasive imaging modality that enables the reconstruction of cross-sectional maps of the scanned object. The process of tomography involves obtaining a series of projections of the object being scanned from various angles and then utilizing reconstruction algorithms to obtain a density field representation of the object based on these projections. Ideally, the acquisition process produces a significant number of uniformly sampled projections across the angular range for high-quality reconstruction. Mathematically, the reconstruction algorithms attempt to solve the inverse problem wherein the unknowns are the pixel/voxel intensities representing the object, while the knowns are the projection values. In some cases such as medical CT scans, reducing the number of projections is desirable to reduce X-ray radiation exposure to the patient. However, the lower dose reconstruction is a challenging problem. To achieve dose reduction, two acquisition scenarios are commonly employed: sparse-view tomography, which uses a limited number of projections, and limited-angle tomography, which samples within a limited angular range. These tomographic reconstruction problems are severely ill-posed and underdetermined. Under such circumstances, the commonly used analytical reconstruction method, Filtered Back-Projection (FBP), becomes obsolete (Zang et al., 2021).

In traditional image reconstruction from limited acquisition, methods have relied on iterative techniques that incorporate prior knowledge and information (Kim et al., 2015; Sagara et al., 2010; Tian et al., 2011). These methods achieve good results but they are computation-

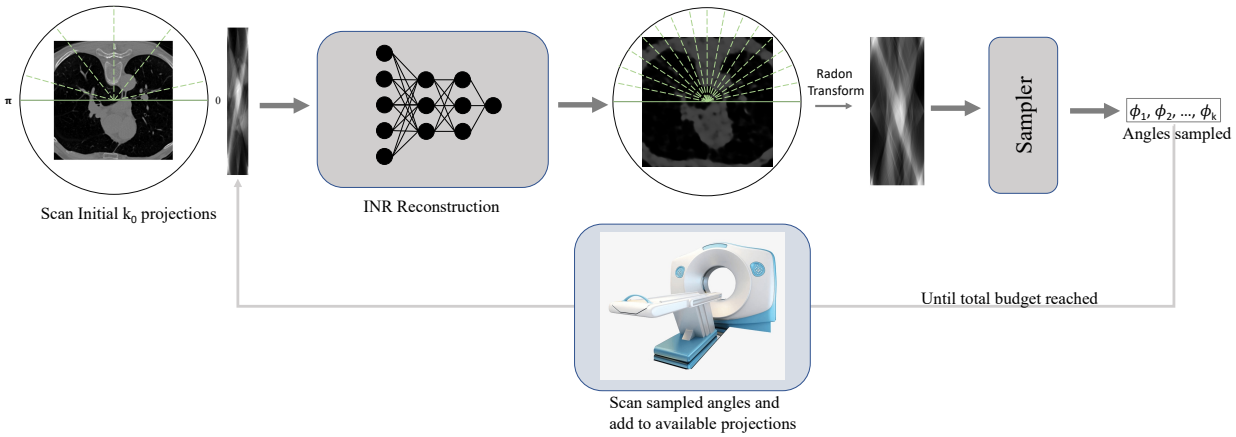


Figure 1: Framework of the active learning method developed. Initially, k_0 projections are sampled at uniformly spaced angles and used to reconstruct a low-quality representation with the INR in Fig. 5. The Radon transform of this reconstruction is input to the sampler, which suggests the next angles to be sampled. The suggested angles are then sampled, and the reconstruction process is repeated using all available angles. The loop is run till the allocated budget is reached.

ally expensive. More recent approaches based on deep learning (DL) aim to directly learn the mapping between low-dose, filtered back-projection (FBP) reconstructed images and high-quality images through paired training (He et al., 2020; Li et al., 2019; Ye et al., 2018). For supervised deep learning-based reconstruction, the availability of training data remains an issue. Thus, a self-supervised framework is desirable (Zang et al., 2021). Among the data-free methods, reconstruction methods with Implicit Neural Representation (INR) as a backbone have shown promising results in both sparse-view and limited-angle scenarios (Song et al., 2023; Zang et al., 2021).

Generally, in a sparse-view setting the projections are sampled uniformly across the desired angular range. Uniform sampling is not adaptive to patients as it does not consider any factor of body representation such as weight, age, and sex (Shen et al., 2020). We hypothesize that reconstruction could benefit from active learning of the projection samples to sample at each progression of reconstruction. We formulate this problem of the active reconstruction as utilizing the sampled projections at each cycle to obtain a low-quality reconstruction and utilizing it to suggest the best projection angles for the next cycle.

In this work, we develop an INR-based reconstructor and investigate the impact of active sampling of projections in a non-uniform manner. The major contributions of the thesis can be summarized as follows:

- We design a self-supervised data-free Implicit Neural Representation based CT reconstruction method that incorporates shape priors and achieves state-of-the-art performance in data-free methods for sparse-view problems.
- We design an active learning framework for CT reconstruction (Fig. 1) and extend the widely used Operator Discretization Library (ODL) to work with non-uniform Radon Transform.
- We investigate the impact of active sampling methods on CT reconstruction and demonstrate that certain projection angles might be more informative to obtain a better reconstruction in a sparse-view scenario.

2. State of the art

2.1. CT Reconstruction

Researchers have approached the CT reconstruction problem with a variety of methods such as initial analytical FBP techniques, iterative reconstruction techniques, a hybrid of these two approaches, and deep learning based data-free and learned reconstruction techniques. A summary of the broad classification of these techniques with advantages and limitations is presented in Table 2.

Analytical methods. The filtered back-projection (FBP) method reconstructs CT slices from projection data by applying a high-pass filter followed by a backward projection step (Willemink and Noël, 2019). The high-pass filter reduces image artifacts and improves the contrast and sharpness of the image. FBP produces images of high diagnostic quality when a large number of projections are available. However, the low-dose reconstruction using FBP results in a significant decrease in image quality with higher image noise and fringing artifacts. While FBP offers the advantage of shorter reconstruction times, the ability to incorporate model and prior information, such as in modeling image noise, is severely limited (Lu et al., 2023).

Iterative and hybrid methods. Iterative reconstruction methods iteratively refine an initial estimate of the image to minimize an objective function that measures

Table 1: Summary of CT Reconstruction Techniques

Reconstruction Technique	Description	Advantages	Limitations
Filtered Back Projection (FBP)	Uses a convolution filter and back projects the measured projections.	Fast reconstruction time.	Limited prior information can be applied. Degraded quality in low dose reconstruction.
Iterative Reconstruction (IR)	Iterates on an objective function to satisfy predefined convergence criteria.	Increased quality images; reduced noise and artifacts; can handle low dose data.	Slow and computationally intensive requires careful optimization and tuning.
Deep Learning-based Reconstruction	Generally data-driven and learned. In some cases used for data-free optimization.	Fast. Can produce high-quality images with reduced noise and artifacts, and can handle low-dose imaging.	Computationally expensive, large data requirement, generalizability, adversarial attacks.

the difference between the measured projection data and forward projected data of the current estimate. IR approaches are particularly well-suited for the ill-posed problem of low-dose CT reconstruction, as they can be combined with regularization and prior terms(such as Total Variation) in an optimization framework (Tian et al., 2011). This helps to improve the stability and quality of the reconstructed images, and can also help to reduce the amount of radiation dose required to obtain diagnostic-quality images. Total Variation (TV) regularization based Reconstruction is often selected as the baseline for Iterative methods (Baguer et al., 2020; Tian et al., 2011; Zang et al., 2021). However, the IR process is computationally expensive and requires significant expertise and optimization to obtain optimal results.

Deep learning methods. Recently, Deep learning (DL) based approaches have been developed which achieve satisfactory performances with fast reconstruction time. These techniques can be classified based on the component of reconstruction they are designed to learn. Post-processing learning employs DL to learn a mapping function between low-quality, sparse view or limited angle FBP reconstruction and high-quality images via paired training. Others are fully learned algorithms that operate on the projection data and do not depend on FBP. A category of these end-to-end methods is learned iterative or unrolling approaches such as learned primal-dual reconstruction which replace the operators in primal-dual optimization with convolutional neural networks (Adler and Öktem, 2018). Fully learned approaches directly learn a mapping from projection data to the image domain through a deep network architecture. In scenarios where training pairs are not available, self-supervised denoising methods such as Noise2Noise (Wu et al., 2020) and data-free or limited data methods such as Deep Image Prior (DIP) have been proposed. In DIP and variants, a U-Net type generative network takes a fixed input of noise and performs the reconstruction iteratively optimizing for the loss of projection data (Baguer et al., 2020). Inspired by NeRF-like methods proposed for view synthesis, co-ordinate based neural representations have also been experimented with for CT reconstruction (Song et al., 2023; Tancik et al., 2021; Zang et al., 2021). A summary of Deep learning-based CT reconstruction methods is presented in Table 2.

2.2. Implicit Neural Representation.

Implicit Neural Representation (INR) is an alternate representation of an image as a continuous function whose input is a pixel coordinate and output is the image intensity at that pixel (Chen et al., 2021; Tancik et al., 2021). This function is parameterized by a multilayer perceptron (MLP). INR has been successfully employed for applications including 3D shape reconstruction (Genova et al., 2019), super-resolution (Chen et al., 2021), novel view synthesis (Mildenhall et al., 2021) and data compression (Dupont et al., 2021). For CT reconstruction with implicit representation, Zang et al. (2021) employ INR wherein an MLP is trained in a self-supervised fashion alongside a geometry refinement module to reconstruct a CT image. Due to promising results with INR, Song et al. (2023) have developed an INR-based framework with test time adaptation which does not require access to training data for hyperparameters tuning. Implicit CT representation and reconstruction thus show promising results significantly outperforming existing approaches on several ill-posed inverse problems of sparse view and limited angle. Thus, we choose to study and develop INR based method as our backbone reconstructor for this project.

2.3. Active Learning.

Active learning is based on the motivation that a machine learning model can achieve greater accuracy if allowed to select the most informative examples from an unlabeled dataset (Gal et al., 2017; Konyushkova et al.,

Table 2: Summary of DL-based CT Reconstruction Techniques. Post-process methods employ Deep Learning to learn a mapping from low-quality FBP representation to high-quality images. Fully learned methods directly learn a mapping from projection data to reconstruction. Learned iterative methods learn the operators such as gradients of iterative reconstruction approaches. INR-based reconstruction methods employ neural representation as an optimization framework.

Method	Strategy	Training type	Summary
Post process	FBPConvNet	Supervised	A U-Net type CNN is trained to reconstruct CT image from sparse view FBP reconstructed image. (Jin et al., 2017)
	RED-CNN	Supervised	U-Net like design with patch based training. (Chen et al., 2017)
	MS-D network	Supervised	Dilated convolutions are used for capturing image features at various scales and layers are densely connected. (Pelt et al., 2018)
	WGAN	Supervised	GAN denoising with Wasserstein distance and perceptual similarity (Yang et al., 2018)
Fully learned	iRadonMAP	Supervised	Neural network with three components: sinogram filtering, back-projection, and refinement. (He et al., 2020)
	Deep Back Projection (DBP)	Supervised	Each view is back-projected separately to form a stack of back projections which is input into a CNN. (Ye et al., 2018)
	iCT-Net	Supervised	Neural network with four components: conversion to dense view sinogram, filtering, backprojection from each view angle, and combination of the partial images. (Li et al., 2019)
	Hierarchical reconstruction	Supervised	A hierarchical framework by casting the original problem as a continuum of intermediate representations. (Fu and De Man, 2019)
Learned Iterative	Learned GD	Supervised	Learning of the gradient for gradient-like iterative reconstruction, making use of prior information, noise model, and a regulariser. (Adler and Öktem, 2017)
	Learned PD	Supervised	Employs CNNs as proximal operators in unrolling a proximal primal-dual optimization method. (Adler and Öktem, 2018)
INR based	DIP-based	Self-supervised	A fixed noise input is passed to U-Net type network to generate the reconstruction by minimizing the loss on projection data. (Baguer et al., 2020)
	Learnit	Pre-training + self-supervised	Coordinate based neural representation with learned initialization. (Tancik et al., 2021)
	IntraTomo	Self-supervised	Coordinate based neural representation with a forward-backward splitting solver-based geometric refinement module. (Zang et al., 2021)
	PINER	Supervised	Two-stage input-adaptation and output-correction framework with implicit neural representation learning. (Song et al., 2023)

2017). The active learner asks queries (unlabeled instances) to be labeled by an oracle. Uncertainty sampling is one of the simplest methods applied to probabilistic learning models (Lewis, 1995). Query-by-committee maintains a committee of models each representing competing hypotheses and trained on the available labeled set. The instance that exhibits the highest level of disagreement among the committee members is considered to be the most informative query (Seung et al., 1992). Another active learning framework is based on expected model change, which prefers the instances that are likely to have the highest impact on model parameters (Settles et al., 2007). In another framework with an estimated error reduction strategy, the expected future error with the labeling of an instance is estimated, and then the instance that minimizes that expectation is selected (Roy and McCallum, 2001). Active learning has seen research related to selecting the most informative examples to label. In medical image analysis, batch mode active learning methods have been proposed for medical image classification (Hoi et al., 2006). More recently, there are also several deep active learning methods for applications such as diabetic retinopathy detection and biomedical image segmentation (Smailagic et al., 2018; Yang et al., 2017). Yang et al. (2017) have shown that state-of-the-art segmentation performance can be achieved by using only 50% of training data.

However, the research on active CT reconstruction remains limited. There are limited prior work with reinforcement Learning (RL) based algorithms where the sampling policy is learned to select the angles (Shen et al., 2020; Wang et al., 2022). However, this framework requires plenty of data and involves a lot of computation and training. Thus, the research question on active CT reconstruction for this project is to investigate the efficacy of simpler active learning techniques on sampling projections. To be more specific, we investigate if certain projections are more important than others if a limited or fixed number of views are to be sampled for reconstruction.

3. Material and methods

In this section, we first establish a background where we introduce the terminologies and notations used and then describe more specific details of our proposed method. Then, we introduce the experimental setup

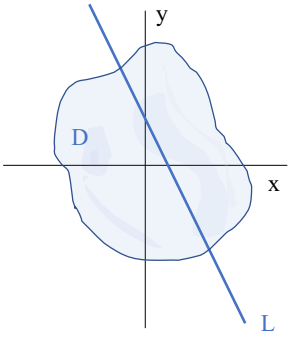


Figure 2: Line L through domain D.

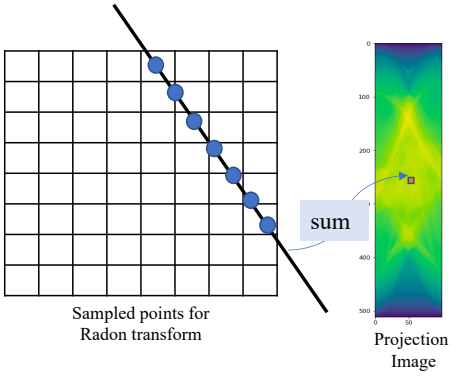


Figure 3: Illustration of Radon transform operator for a ray passing through 50 degree angle.

which includes the datasets, the evaluation metrics, and the implementation details.

3.1. Background

3.1.1. Radon Transform

Let (x, y) be the coordinates of points in the plane, and f be an arbitrary function defined on some domain D of $\mathbb{R}^2$. For any line L in the plane, then the mapping defined by the projection or line integral of f along all possible lines L is the 2D Radon transform R of f provided the integral exists (Deans, 2007). Mathematically,

$$\check{f} = Rf = \int_L f(x, y)ds, \tag{1}$$

where ds is an increment of length along L. The domain D could be the entire plane or some region of the plane as shown in Fig. 2.

In 2D images, the Radon transform on an image for a set of angles is the sum of the intensities of the pixels in each direction. Fig. 3 depicts a ray passing at a projection angle of 50 degree, and the blue dots represent the coordinates used to calculate the line integral.

3.1.2. CT Reconstruction

In CT reconstruction, the projection data represents the line integral or the attenuation of the x-ray when

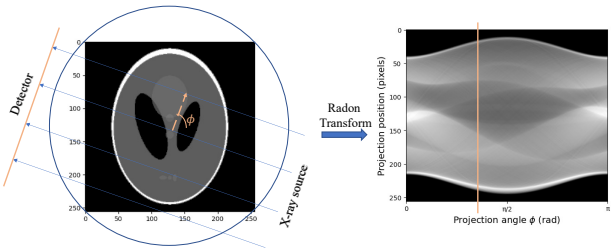


Figure 4: Parallel beam geometry with x-ray beam projected at evenly spaced 300 angles ϕ from 0 to π to obtain sinogram on the right.

passing through a body, which is given by 2D Radon transform. In this work, we consider a parallel beam geometry where the x-ray beam is collimated to form a parallel beam as illustrated in Fig. 4. In the first diagram, four dotted parallel lines represent the rays of the x-ray beam and the yellow line parallel to it represents the detector. The yellow line in the second diagram in Fig. 4 represents the projection that corresponds to the projection angle in the first diagram. Here, projection angle ϕ is the angle between the x-axis and the line of projection. A sinogram is obtained by taking a series of projections of an object at a series of projection angles. The goal of CT reconstruction is to perform an inverse Radon transform to estimate an image from a sinogram. However, this inverse problem is highly ill-posed when the reconstruction is to be done from a sparse view of projections.

The problem can be formulated by Radon operator $A : X \rightarrow Y$ from space X to Y and the measured noisy projection data:

$$y^{\delta} = Ax^{\dagger} + \tau. \tag{2}$$

where y^{δ} is the acquired noisy projection, $x^{\dagger}$ is the true solution and $\tau \leq \delta$ is the noise in the projection. The objective of CT reconstruction is to obtain an approximate $\hat{x}$ for $x^{\dagger}$.

3.1.3. Implicit Neural Representation (INR)

In INRs, a CT image is represented as a function whose input is pixel coordinate $\mathbf{p}_i = (x_i, y_i)$ and output is the gray-scale intensity value at $\mathbf{p}_i$. Let N be the number of pixels that discretize the image space. This continuous function is approximated with an MLP network $F_{\Theta} : p \rightarrow I$ and the weights Θ are optimized to map the pixel coordinates to intensity value.

$$F_{\Theta}(\mathbf{p}_i) = \mathbf{I}_i, \quad i = 1..N \tag{1}$$

We describe the INR design used in more detail and also present an overview diagram in Fig. 5.

3.1.4. Gaussian random Fourier feature Encoding.

Positional encoding has been shown to improve the high-frequency details in applications such as neural

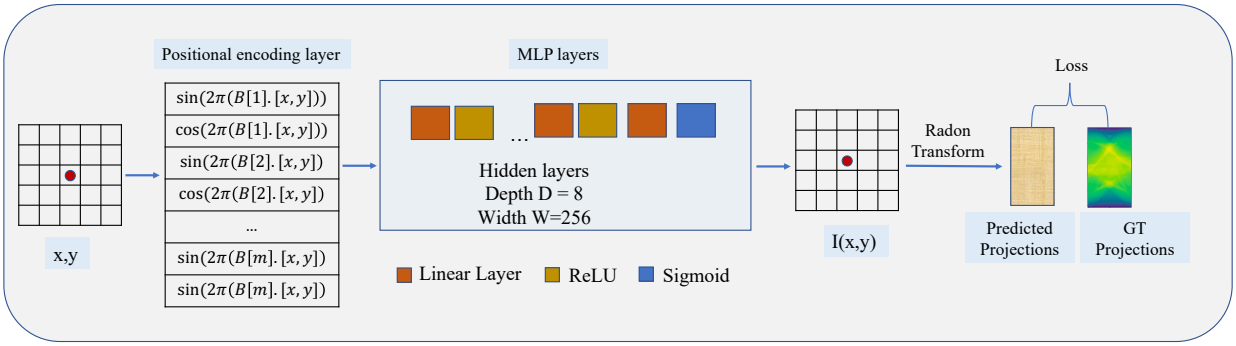


Figure 5: Architecture used for the implicit neural representation and reconstruction. The input of the network is pixel coordinate and the output is intensity at that pixel coordinate. Positional encoding (random Fourier feature encoding) is applied to the pixel coordinate before inputting to the MLP. The intensity values represent the image on which Radon Transform is applied to obtain projections. The loss is calculated on the projections and backpropagated.

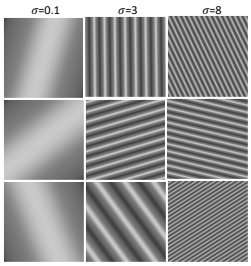


Figure 6: Gaussian Fourier features at different σ.

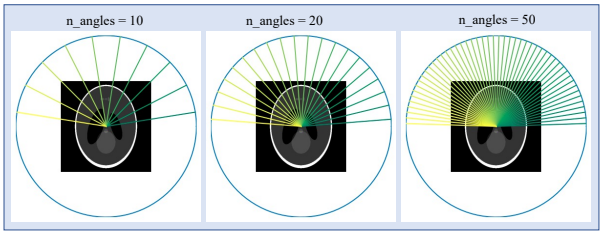


Figure 7: Illustration of projection angles sampled by USampler.

scene rendering, low-dimensional regression, and also CT reconstruction (Mildenhall et al., 2021; Tancik et al., 2020; Zang et al., 2021). This encoding is formulated as a mapping of a low-dimensional point to a high-dimensional space with a set of sinusoidal features, also called Gaussian random Fourier features. Formally, the mapping we use is given by the encoding function $\gamma(.)$, which is applied elementwise.

$$\gamma(\mathbf{v}) = [\cos 2\pi \mathbf{B}\mathbf{p_i}, \sin 2\pi \mathbf{B}\mathbf{p_i}]^\top \quad (3)$$

where $B \in R^{m\times 2}$ is sampled from Gaussian distribution $\mathcal{N}(0, \sigma^2)$, and m is the encoding size. We visualize the encoded pixel space visualized at the resolution $(256, 256)$ in Fig. 6. We can see that the high value of σ represents higher frequencies and the low value represents lower frequencies. Thus, the selection of σ depends on the application. We present the ablation on the choice of this encoding later in the section 4.4.2.

3.1.5. Active learning

In active learning literature, random sampling is a common method chosen for comparison with the developed sampling method. In general, it is not straightforward to apply the general active learning methods to the problem of Active CT Reconstruction. We explain the framework designed in Section 3.2.2. Here we explain the sampling methods which are commonly used as a comparison standard for active learning problems we consider as our baseline.

Uniform sampler (US). Uniform Sampler is the simplest possible case of the sampler in which evenly-spaced projections are sampled from the angular range of 0 to π. To illustrate, we present examples of angular sampling with 10 and 20 angles in Fig. 7.

Core-set sampler (CSS). CSS is a k-center-greedy algorithm that minimizes the maximum distance of any point to a center (Sener and Savarese, 2017). In our case, at each cycle, the greedy algorithm chooses the k projections which are at the furthest distance from the current center of projections. We use Euclidean distance as the distance metric for the Core-set sampler.

3.2. Proposed method

3.2.1. INR Design

Here we describe the more specific components of our INR design: MLP Layers, Radon Transform specifics, and Loss functions.

MLP Layers. The MLP consists of 8 fully-connected layers with 256 neurons in each layer, except the last layer which has one neuron. Each linear layer is followed by a ReLU activation while the final layer goes through a Sigmoid activation. With the representation presented above, we can input the pixel coordinates in a desired discretized space to the MLP to obtain an image output.

Loss functions and regularization. After applying the Radon Transform operator on the output of the MLP, predicted projections are obtained. Loss is computed between the predicted projections and the ground truth projections. We experiment with L2-loss and Poisson loss (details provided in Section 3.3.3). We add Total Variation (TV) regularization term to the loss, whose weight is tuned for each dataset. TV regularization is shown to recover sharp edges while it looks for a solution with minimal total variation (Strong and Chan, 2003). The TV regularization term for an intermediate output x is:

$$TV(x) = \sum_{i,j} |x_{i+1,j} - x_{i,j}| + |x_{i,j+1} - x_{i,j}| \tag{4}$$

Early Stopping. Optimization-based algorithms require stopping criteria to avoid over-fitting to noise. We use early stopping based on loss with patience of 2000 iterations.

3.2.2. Active Learning

Active learning literature is focused on the sampling of the most informative examples to label to improve model performance. It is not straightforward to apply these developed concepts to CT reconstruction because they require a pool of observations from which suitable samples need to be labeled. Nevertheless, we investigate some active learning methods adapting them to the CT reconstruction problem. We assume that the projections are available to investigate for evidence that active learning is suitable for CT reconstruction.

We structure the active learning framework for CT reconstruction as learning a discrete selection problem. The active sampler should learn a sampling matrix P to sample the projections from $A\ x^{\dagger}$ (ground truth projections) to minimize the projection error and regularization terms as mathematically formulated in Eq. 5.

$$\min_{x,P} \frac{1}{1} \| P\ A\ x - y^{\delta 2} \| + \alpha TV(x) \tag{5}$$

where α is the regularization weight.

Active reconstruction thus has two main components and we alternate between the following two steps until desired criteria are met. The iterative scheme is depicted in Fig. 8.

1. Reconstruction: At any cycle k, the task of the reconstructor is to reconstruct image given the obtained projections at that cycle. We have designed this step as optimizing the INR MLP described in 3.1.3 to better fit the projections available at each cycle. Initially, k_0 projections are uniformly sampled so that the reconstructor learns a low-quality representation to be passed to the sampler which would then suggest the best projections to sample from based on the current reconstruction. In terms of active learning terminologies, k_0 is analogous to *initialbudget*.

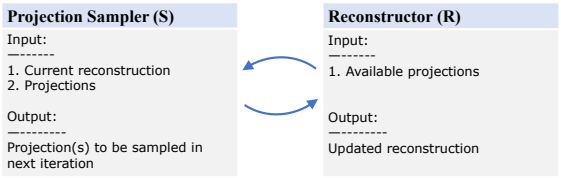


Figure 8: Iterative scheme of simultaneous reconstruction and sampling. The Sampler takes in current reconstruction and projections as input and the Reconstructor takes in available projections to reconstruct.

2. Sampling: At each cycle n, the sampler S samples k projections corresponding to angles ϕ_k. Here, k can be considered as the *budget* in active learning context. The sampler is expected to sample the projections that would lead to the best possible reconstruction, given the current reconstruction. If the reconstruction at the end of cycle n is x_n and the sampling matrix is P_n, the sampled projections at cycle $n+1$ can be represented as:

$$p_{n+1} = S(x_n, P_n) \tag{6}$$

Since we do not have access to the ground truth reconstruction and only have access to the projection data, one way to sample the projections is to choose the projections for which the current error is highest. It can be expected that sampling projections with the highest error might lead to low projection error and possibly higher quality reconstruction. We describe the methods investigated for sampling in the following section.

Projection error based sampler (PES). The hypothesis behind sampling a projection with the highest error at any cycle is that sampling such a projection should reduce the projection error after undergoing the next reconstruction cycle, which in turn should improve the reconstruction quality. Thus, at each cycle, we calculate the current projections (discretized at 1000 angles in the range 0 to π), calculate the Euclidean distance to the ground truth projections, and sample the projection with the highest distance.
Let $\hat{y} = Ax_k$ be the predicted projections at cycle k, and y^{δ} be the ground truth projections. The selection criteria of PES can be mathematically formulated as:

$$\min_{\phi} \sqrt{\sum_{\phi \in Phi} (\hat{y}_{\phi} - y^{\delta}{}_{\phi})^2} \tag{7}$$

where j represents the discretized space from which sampling is done.

Projection error based sampler constrained on angle distribution (C-PES). In some poorer cases of reconstruction with sampling by PES, we observe that the

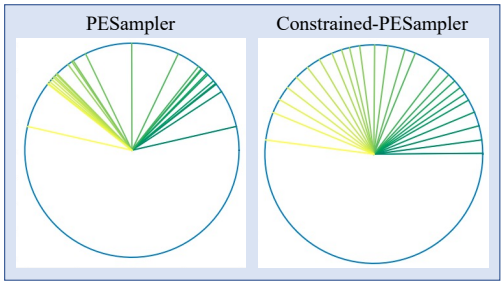


Figure 9: An example of the angles sampled by PESampler vs. Constrained-PESampler.

projection angles are concentrated around one angular region (first diagram in Fig. 9). This implies that the projection error alone might not be a good criterion for sampling and some uniformity is also required so that the sampled projections contain enough information in all directions.

In order to enhance the PES, we introduce a simple constraint. During each sampling cycle, we ensure that the angle sampled is no less than 49.5% of the current division of the projection angular region. For instance, if 10 projection angles have been sampled in a given cycle and an equal division is assumed, each angle would differ by 0.314 radians. With the added constraint, the chosen angle must be at least 49.5% of half the value of 0.314 radians.

3.2.3. Optimization Strategy

The Reconstructor operates on the available projection data through the INR fitting strategy with a loss on projection data. In each cycle of active sampling, the Sampler samples k projections. While the Reconstructor aims to minimize the projection loss, the Sampler suggests the best angles to sample from. In this work, we assume the projections are available to demonstrate that CT reconstruction does benefit from active learning. The Sampler is a defined method and not learned. Thus, INR fitting is updated each cycle minimizing the loss on all the sampled projections.

3.3. Experimental setup

3.3.1. Datasets

We run our experiments on the following datasets:

1. Shepp-Logan phantoms: Shepp-Logan phantom was first created by Shepp and Logan (1974) and has been used as a standard test image by CT reconstruction research. We use 50 randomly generated 256x256 pixel Shepp-Logan phantoms as ground truth and we generate the projections by applying the Radon operator with a parallel beam geometry with 256 rays. We experiment with different numbers of available projections in a sparse-view setup. Additionally, we add a white noise with a standard deviation of 2.5% of the mean absolute value of the projection data to the projection data. 5 separate samples are used for tuning hyperparameters which are not included in the test set. We show some Shepp-Logan phantoms in Fig. 10.

Algorithm 1 Active reconstruction framework.
reconstruct = a method that fits the given *projections* into an MLP.
k_0, k, *n_angles* = initial budget, budget, maximum budget

Require: *projections*, *iterations*, *f_iterations*, k, n, k_0
Require: *reconstruct*
  initial_samples ← *uniform_partition*(0, π, k_0)
  sampler ← *sampling_strategy*
  selected ← *initial_samples*
  reconstruct(*selected*)
  for each cycle in range(*n_angles* − k_0) **do**
    new_sel ← *sampler.sample*(*gt_proj*, *pred_proj*, k)
    selected ← *selected* + *new_sel*
    reconstruct(*selected*)
  end for
  best_output ← *reconstruct*(*selected*)

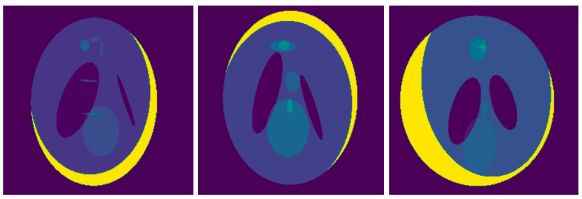


Figure 10: Some samples of randomly generated Shepp-Logan phantoms.

2. LoDoPaB-CT dataset: Low-dose parallel beam (LoDoPaB) CT dataset is a benchmark dataset for low-dose ct reconstruction designed by Leuschner et al. (2021). In this dataset, the ground truth is composed of 362 × 362-pixel human chest CT reconstructions. The projections are simulated with a simple parallel beam geometry with 1000 angles and 513 projection beams. We sub-sample the angles from the available 1000 to simulate even sparser reconstruction scenarios. We use 50 images as a test set and 5 for tuning hyperparameters. We show some LoDoPaB-CT sampled in Fig. 11.

3.3.2. Evaluation metrics

We focus on two evaluation metrics which are widely used to evaluate CT reconstruction performance:

Peak signal-to-noise ratio (PSNR)

PSNR is the ratio between the maximum possible value of an image and the power of noise that affects the quality of its representation. For an image f and its approximation g of size m x n, PSNR in decibels (dB) is defined as:

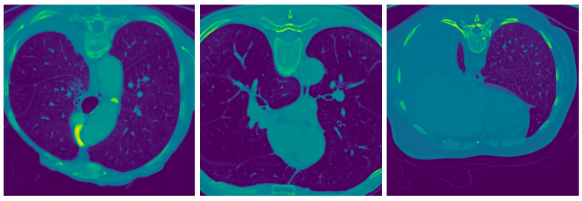


Figure 11: Some samples of LoDoPaB-CT dataset.

$$PSNR = 20log_10(\frac{MAX_f}{\sqrt{MSE}}) \quad (8)$$

where MSE is the Mean Squared Error calculated as:

$$MSE = \frac{1}{mn}\sum_{0}^{m-1}\sum_{0}^{n-1}||f(i,j) - g(i,j)||^2 \quad (9)$$

The higher the PSNR, the better the reconstruction quality.

Structural Similarity Index (SSIM)

SSIM is formulated by representing image distortion as a composite of three contributing factors, loss of correlation, luminance distortion, and contrast distortion (Hore and Ziou, 2010).

$$SSIM(f,g) = l(f,g)c(f,g)s(f,g) \quad (10)$$

where $l(f,g) = \frac{2\mu_f\mu_g+C_1}{\mu_f^2+\mu_g^2+C_1}$, $c(f,g) = \frac{2\sigma_f\sigma_g+C_2}{\sigma_f^2+\sigma_g^2+C_2}$, and $s(f,g) = \frac{\sigma_{fg}+C_3}{\sigma_f\sigma_g+C_3}$

$l(f,g)$ is luminance comparison function where μ_f and μ_g denote mean luminance. $c(f,g)$ is a contrast comparison function where contrast is measured by the standard deviation σ. $s(f,g)$ measures the correlation between the images and σ_{fg} is the covariance between f and g. Constants $C1$, $C2$, and $C3$ are added to avoid zero denominators.

SSIM index values lie in [0,1] where 0 implies no correlation and 1 implies f and g are the same image.

3.3.3. Implementation Details

We implement all the methods in PyTorch. Active sampling requires that the Radon transform should be able to compute projections at non-uniformly sampled angles. Thus, we extend the Operator Discretization Library (ODL) for non-uniformly sampled Radon Transform operations to compute projections for any set of input projection angles. We use the Dival Library to access the LoDoPaB-CT dataset and extend it for Phantom dataset. We use L2 and Poisson loss for the Phantom dataset and LoDoPaB-CT dataset respectively. We use Adam Optimizer for all experiments and the learning rate starts from 0.001 for the Phantom dataset and 0.0005 for the LoDoPaB-CT dataset. We use a learning rate scheduler to reduce the learning rate on plateau with a patience of 1500 iterations. We set the budget k to 1, and initial budget k_0 to half of the total budget. After sampling new projection in each cycle, we run the optimization for 100 iterations. We evaluate the Phantom Dataset at 256×256 resolution and the LoDoPaB-CT Dataset at 362×362 resolution. All experiments are performed on NVIDIA V100 GPU. A summary of all hyperparameters is presented in Appendix A.

Table 3: PSNR/SSIM summary on Phantom (PD) and LoDoPaB-CT (LD) dataset at different angles.

PD	8	10	15	20
FBP	8.50/.102	9.73/.130	12.18/.199	13.72/.242
TV	19.48/.536	22.62/.64	29.64/.870	31.65/.922
DIP	**28.41/.916**	30.40/.949	31.65/.960	32.14/.965
INR	28.39/.901	**33.05/.951**	**35.59/.972**	**36.31/.976**
LD	25	50	100	200
FBP	10.45/.107	18.45/.141	22.38/.241	28.38/.649
TV	27.64/.695	29.03/.732	30.06/.757	30.86/.776
DIP	28.34/**.719**	29.82/**.756**	31.19/.773	31.89/**.800**
INR	**28.94/.719**	**30.44/.756**	**31.32/.777**	**32.21/.800**

4. Results

4.1. INR base method vs. baseline methods

We compare the reconstruction quality of the developed base INR model with FBP, TV-based reconstruction, and Deep Image Prior. For FBP, we use a simple ramp filter. For Iterative and DIP reconstruction, we use the implementation by Baguer et al. (2020). In Table 3, Fig. 12, Fig. 13, we show that INR has the best performance among the compared self-supervised methods both quantitatively and qualitatively for both datasets. DIP and INR are far better than FBP and TV. However, the proposed INR outperforms DIP at all evaluated angles for the LoDoPaB-CT dataset.

4.2. Active learning

We choose the INR as a base reconstructor and evaluate the reconstruction results using different greedy sampling techniques: PES and C-PES. We report the PSNR and SSIM of the sampling methods, and US and CSS in Table 4 for the Phantom dataset and Table 5 for LoDoPaB-CT dataset. For the Phantom dataset, C-PES achieves better reconstruction in all angles except when *n_angles* is 20, where the US and CPES attain similar metrics. Similarly, for the LoDoPaB-CT dataset, we observe improved PSNR and SSIM across all evaluated angles. However, we notice that the improvement over US diminishes as more angles are sampled. In the case of the Phantom dataset, the PSNR improvement is 1.48dB for 8 views, while for 20 views, the average PSNR on this dataset is equal. We observe a similar trend for the LoDoPaD-CT dataset. The improvement in PSNR is 0.93DB for 25 views, 0.32dB for 50 views, 0.16dB for 100 views, and 0.11dB for 200 views.

Table 4: Comparison of reconstruction on Phantom dataset with different sampling techniques on the INR base reconstructor.

	n_angles=8		*n_angles*=10		*n_angles*=15		*n_angles*=20	
Sampler	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
US	28.39	.901	33.05	.951	35.59	.972	36.31	.976
PES	29.40	.921	32.47	.958	34.52	.971	35.74	.976
CSS	27.97	.897	30.92	.939	34.31	.966	31.63	.779
CPES	**29.87**	**.927**	**33.34**	**.958**	**35.67**	**.973**	**36.31**	**.976**

Table 5: Comparison of reconstruction on LoDoPaB-CT dataset with different sampling techniques on the INR base reconstructor.

	n_angles=25		*n_angles*=50		*n_angles*=100		*n_angles*=200	
Sampler	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
US	28.53	.712	30.12	.750	31.16	.773	32.10	.787
PES	28.01	.699	29.37	.730	30.56	.756	31.63	.780
CSS	28.33	.708	29.68	.737	30.74	.761	31.63	.779
CPES	**28.94**	**.719**	**30.44**	**.756**	**31.32**	**.777**	**32.21**	**.800**

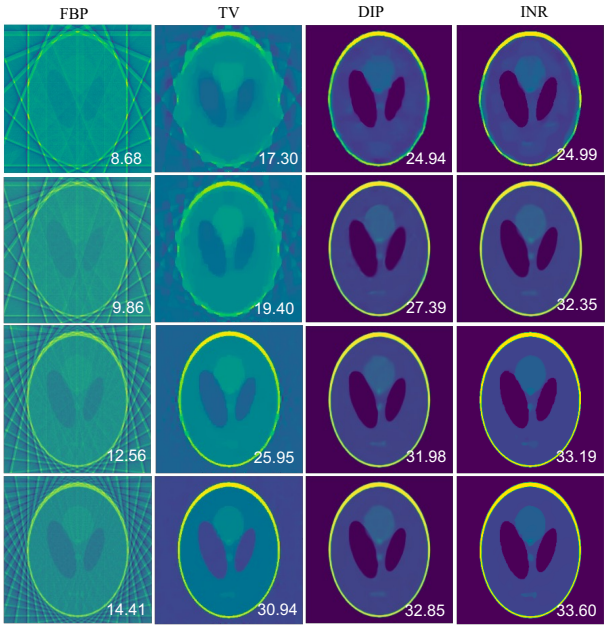


Figure 12: Reconstruction at different numbers of projection angles 8, 10, 15, and 20 from top to bottom on Shepp-Logan phantom. The number at the bottom right refers to the PSNR value.

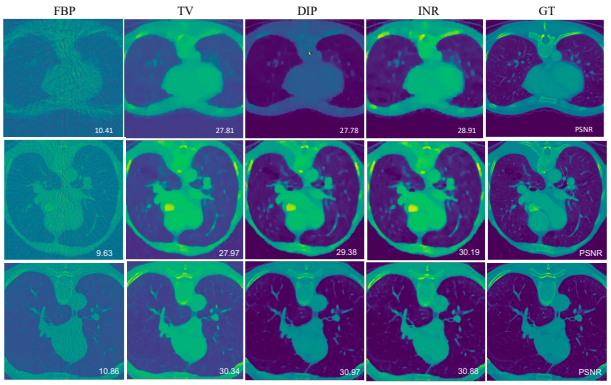


Figure 13: Reconstruction comparison of baseline methods on LoDoPaB-CT dataset. The number at the bottom right refers to the PSNR value.

To present the qualitative results, we illustrate four examples of reconstruction using only 8 projections for the Phantom dataset in Fig. 14. Despite the limited number of angles, the C-PES method consistently exhibits excellent reconstruction results, wherein small ellipses are distinctly visible in comparison to US or CSS approaches. In certain cases, the US method produces a reconstruction that is slightly better, as observed in row 2. However, there are cases where US exhibits comparatively poorer reconstruction results, characterized by dissolved major structures or an unsmooth texture, as seen in row 1.

Similarly, we present the reconstruction results on LoDoPaB-CT dataset at different angles in Fig 15. As this is a more complex dataset with finer details, we also present a zoomed-in snapshot of the reconstructed image in the second and fifth rows. At 25 angles (row 1, example 1), we observe that the C-PES retains finer details more effectively compared to US or CSS. We observe a similar trend with 50 angles (row 1, example 2), where the reconstruction demonstrates slight improvements in capturing details. We also present a scenario (row 3, example 1) where C-PES exhibits a slightly lower PSNR (0.16dB) when compared to US. However, upon visual inspection, the reconstructions appear similar. Furthermore, in row 3, example 2, we observe that C-PES achieves slightly superior reconstruction, particularly in preserving smaller structures.

4.3. Analysis of Sampled angles

We illustrate the angles sampled by the sampling processes in Fig. 14 (rows 2 and 4) and Fig. 15 (rows 3 and 6). We notice that the angles that lead to better reconstruction are not always uniformly spaced. US has sig-

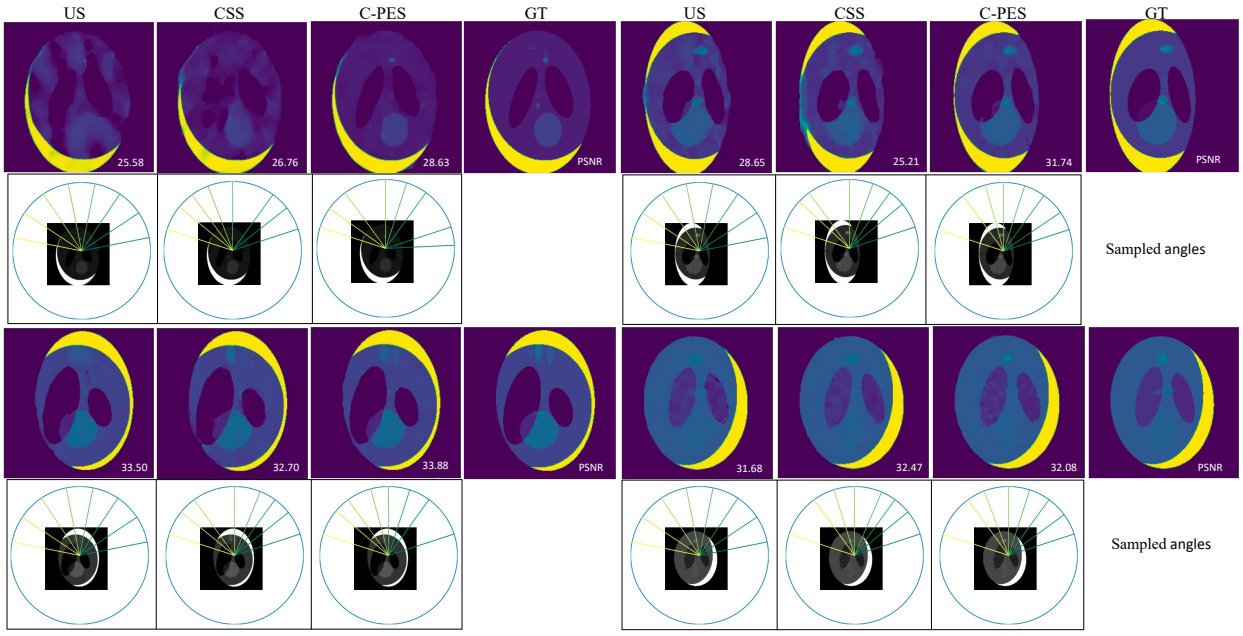


Figure 14: Reconstruction with 8 projection angles on Phantom dataset. The second and fifth rows visualize the zoomed-in image. The number at the bottom right refers to the PSNR value.

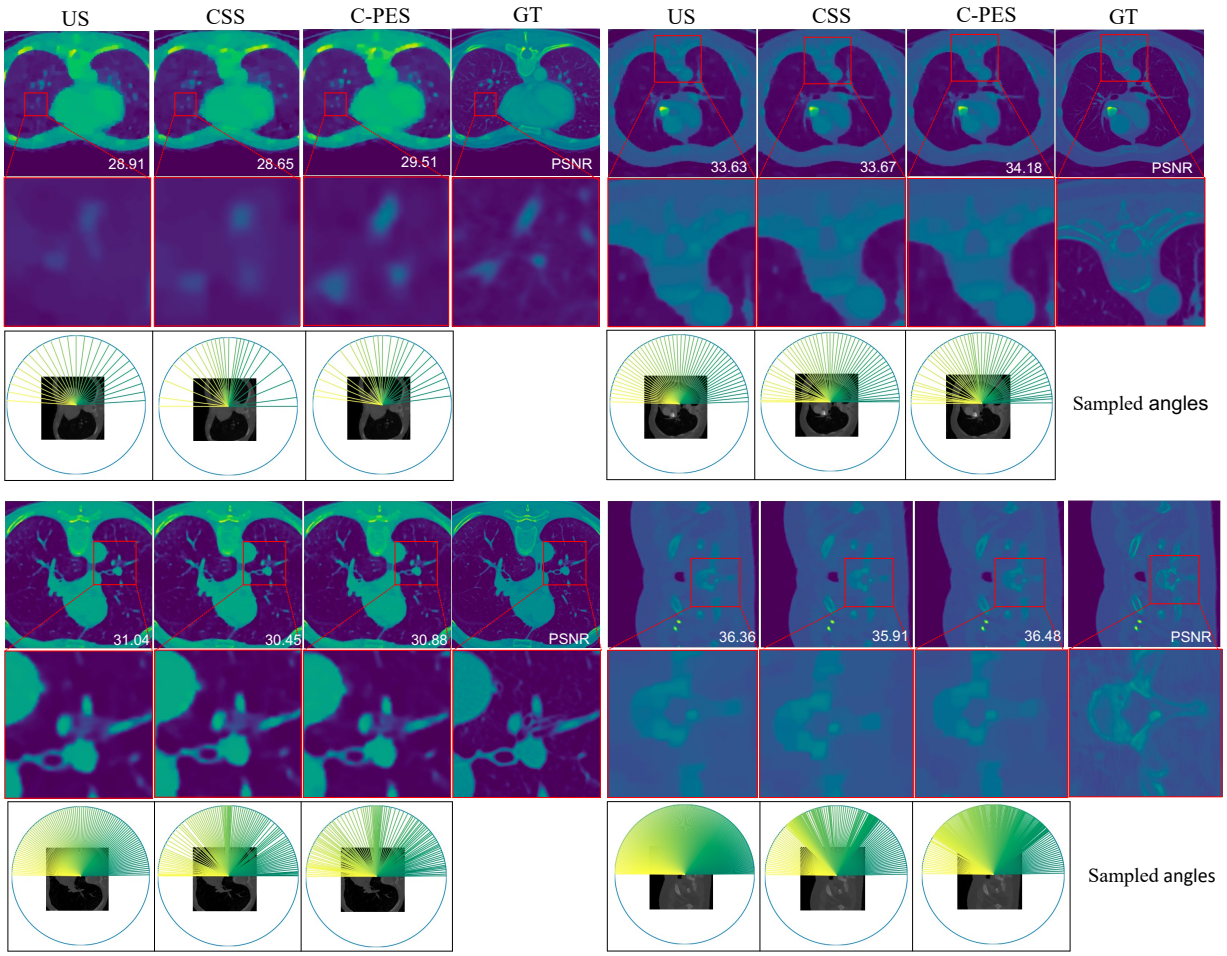


Figure 15: Reconstruction at different numbers of projection angles 25, 50, 100, and 200 on LoDoPaB-CT dataset. The second and fifth rows visualize the zoomed-in image. The number at the bottom right refers to the PSNR value.

nificantly poorer performance for certain Phantom examples (row 1 in Fig. 14). When the angles are analysed, we observe that the angles sampled by C-PES corresponding to these reconstructions are different compared to US. However, in row 2, example 2, we can observe that the reconstruction by US is close to C-PES. Upon inspection of the angles sampled by C-PES, we observe that the angles are close to uniform sampling. Similarly, for LoDoPaB-CT dataset, we can observe that the best reconstruction obtained by C-PES does not correspond to US angles.

4.4. Ablation studies

4.4.1. Radon Transform Implementation

We can utilize the continuous representation of an image to accurately compute Radon transform, which requires interpolation when being computed at different projection angles. In Fig. 3, we visualize an example of Radon transform for a ray at a projection angle of 50 degrees. For the implementation of Radon transform, we explored two strategies: (i) directly applying Radon operator on the image rendered by the INR, or (ii) inputting the points sampled to calculate the integral and obtain the intensity value from the network. In Fig. 3, we illustrate the process of calculating the projection value for a ray. In the case of equally spaced parallel beam acquisition, we can notice that the sampled points are not always located at the center of a pixel. Consequently, the projections computed using the first strategy may lack precision. We compared the reconstruction quality between computing the predicted Radon transform from the rendered image and predicting the intensities at the required continuous pixel positions and summing them. Although the second implementation resulted in slightly better quality reconstruction (0.91 dB PSNR improvement), we chose to use the first implementation due to its faster reconstruction time (2.1 times faster).

4.4.2. Hyperparameter choices

We discuss the choice of hyperparameters involved in the developed methods: σ, m in Gaussian Fourier feature encoding. In Fig. 17, we present the plot between σ and m with evaluation metrics, justifying our choice for $\sigma = 1, m = 256$ for the Phantom dataset and $\sigma = 4.5, m = 256$ for LoDoPaD-CT dataset. In Fig. 18, we show reconstruction with varying σ to study the impact of this parameter. We observe that lower value of σ leads to smooth reconstruction and a higher value leads to granular, which can also be anticipated in Fig. 6 where we present the Gaussian Fourier features. The TV regularization weights (α) are taken from Baguer et al. (2020). The impact of TV weights is presented in Fig. **??**. We can see that a lower value gives a smooth reconstruction and a granular reconstruction. Thus, it needs to be tuned according to the application.

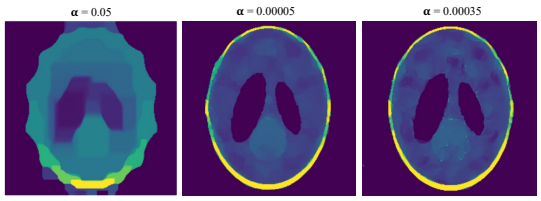


Figure 16: Visualization of reconstruction at different TV Regularization weight (α).

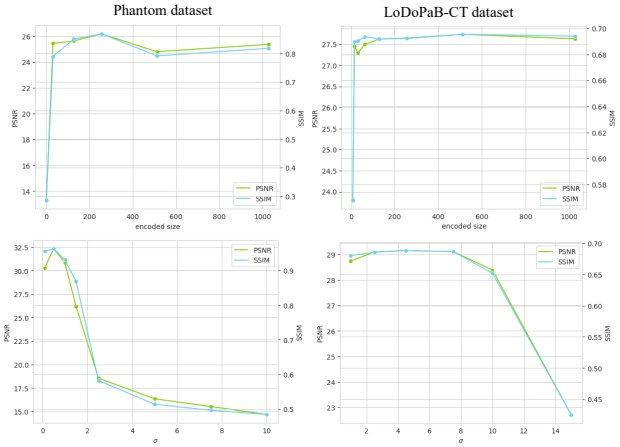


Figure 17: Evaluation metrics for positional encoding parameters: m (encoded size), and σ (standard deviation of the Gaussian distribution from which random Fourier feature is generated).

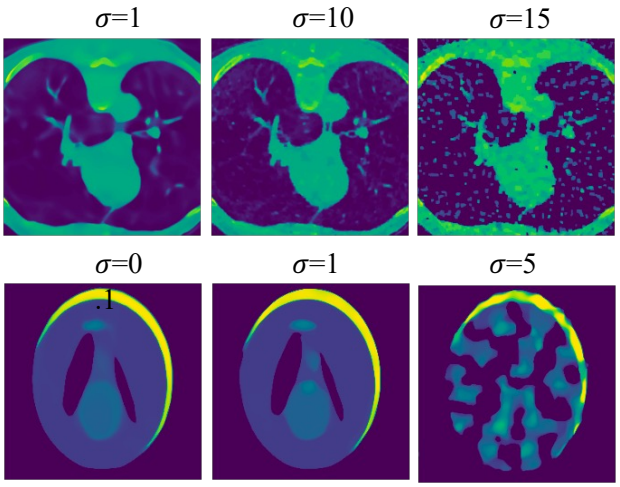


Figure 18: Visualization of reconstruction at different σ of positional encoding (Gaussian Fourier features).

5. Discussion

5.1. Insights

From the results presented above, we discuss that neural representations have the capability to produce qualitatively better reconstructions, with a significant increase in the PSNR and the SSIM values, particularly in sparse scenarios. The developed INR approach already outperforms the data-free baseline DIP, however,

learning the optimal initialization weights as proposed in Tancik et al. (2021) should lead to more efficient reconstruction.

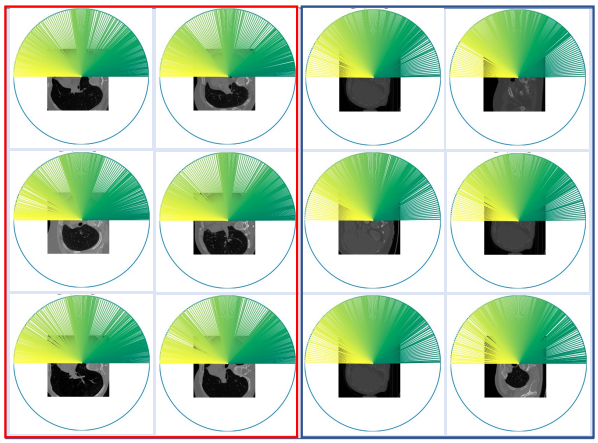


Figure 19: Visualization of the sampled angles for different slices of LoDoPaD-CT dataset.

We gain valuable insights into the selection of projection angles for achieving higher-quality CT reconstruction. It is evident from our results that non-uniform sampling significantly benefits the reconstruction process. Furthermore, in Fig. 19, we visualize the sampled angles that lead to higher PSNR. We observe that similar slices exhibit a similar distribution of angles, as indicated by the grouping of these slices into red and blue boxes. This finding suggests a correlation between the set of angles and the specific anatomical structures being reconstructed. Based on this observation, we can hypothesize that incorporating prior knowledge about the anatomy into the selection of projection angles, in the form of a preset angle configuration, might be beneficial for sparse-view CT reconstruction.

5.2. Limitations and Future Work

A key limitation of our study is the assumption of access to a pool of projections and using those to draw conclusions regarding the effectiveness of Active Sampling of angles. However, in real-time CT acquisition scenarios, such a pool of projections is not readily available. If a pool of projections is available, it means that the patient has already undergone radiation, which drives away from our motivation. Consequently, future research should focus on addressing this limitation by incorporating meta-knowledge learning to dynamically select the most informative angles during each cycle of active sampling. This approach would enable us to adaptively acquire projections without the need for a pre-existing pool, aligning with the practical constraints of low-dose CT imaging. By exploring the integration of meta-knowledge into the active sampling process, we can enhance the effectiveness and applicability of our proposed reconstruction framework in real-world clinical settings.

6. Conclusions

In this work, we design and investigate the efficacy of Implicit Neural Representation for CT reconstruction with low-dose CT reconstruction and the role of active sampling in improving reconstruction. First, we propose an INR network and training strategy, and then an active learning framework capable of handling non-uniform projection sampling to select the best projections. Through extensive experiments on both synthetic phantom (Shepp-Logan) and real patient (LoDoPaB-CT) datasets, we have demonstrated the benefits of active sampling in the context of sparse view CT reconstruction. However, our current framework does not involve the learning of meta-knowledge for selecting the optimal sampling strategy. While our proposed active learning framework shows promising results, it is not straightforward to apply in real-time scenario. To address this limitation, we intend to explore the possibilities of meta-learning in the future, aiming to develop a more robust and adaptable sampling strategy that can generalize across different datasets and imaging conditions. Overall, our findings highlight the potential of INR-based reconstruction methods and the significance of active sampling in CT imaging. By continuing to refine and expand upon these methodologies, we anticipate further advancements in low-dose CT reconstruction techniques, ultimately contributing to improved diagnostic accuracy and reduced patient radiation exposure.

Acknowledgments

I would like to extend my gratitude to the European Union and the MAIA team for providing funding and support throughout my master's program. Additionally, I would like to express my sincere appreciation to my supervisors for their invaluable guidance, expertise, and insightful feedback during the thesis development. I am also grateful to the WEISS team for hosting me and providing a conducive research environment.

References

Adler, J., Öktem, O., 2017. Solving ill-posed inverse problems using iterative deep neural networks. Inverse Problems 33, 124007.

Adler, J., Öktem, O., 2018. Learned primal-dual reconstruction. IEEE transactions on medical imaging 37, 1322–1332.

Baguer, D.O., Leuschner, J., Schmidt, M., 2020. Computed tomography reconstruction using deep image prior and learned reconstruction methods. Inverse Problems 36, 094004.

Chen, H., Zhang, Y., Kalra, M.K., Lin, F., Chen, Y., Liao, P., Zhou, J., Wang, G., 2017. Low-dose ct with a residual encoder-decoder convolutional neural network. IEEE transactions on medical imaging 36, 2524–2535.

Chen, Y., Liu, S., Wang, X., 2021. Learning continuous image representation with local implicit image function, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8628–8638.

Deans, S.R., 2007. The Radon transform and some of its applications. Courier Corporation.

Dupont, E., Goliński, A., Alizadeh, M., Teh, Y.W., Doucet, A., 2021. Coin: Compression with implicit neural representations. arXiv preprint arXiv:2103.03123 .

Fu, L., De Man, B., 2019. A hierarchical approach to deep learning and its application to tomographic reconstruction, in: 15th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine, SPIE. p. 1107202.

Gal, Y., Islam, R., Ghahramani, Z., 2017. Deep bayesian active learning with image data, in: International conference on machine learning, PMLR. pp. 1183–1192.

Genova, K., Cole, F., Vlasic, D., Sarna, A., Freeman, W.T., Funkhouser, T., 2019. Learning shape templates with structured implicit functions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7154–7164.

He, J., Wang, Y., Ma, J., 2020. Radon inversion via deep learning. IEEE transactions on medical imaging 39, 2076–2087.

Hoi, S.C., Jin, R., Zhu, J., Lyu, M.R., 2006. Batch mode active learning and its application to medical image classification, in: Proceedings of the 23rd international conference on Machine learning, pp. 417–424.

Hore, A., Ziou, D., 2010. Image quality metrics: Psnr vs. ssim, in: 2010 20th international conference on pattern recognition, IEEE. pp. 2366–2369.

Jin, K.H., McCann, M.T., Froustey, E., Unser, M., 2017. Deep convolutional neural network for inverse problems in imaging. IEEE Transactions on Image Processing 26, 4509–4522.

Kim, Y., Kim, Y.K., Lee, B.E., Lee, S.J., Ryu, Y.J., Lee, J.H., Chang, J.H., 2015. Ultra-low-dose ct of the thorax using iterative reconstruction: evaluation of image quality and radiation dose reduction. American journal of roentgenology 204, 1197–1202.

Konyushkova, K., Sznitman, R., Fua, P., 2017. Learning active learning from data. Advances in neural information processing systems 30.

Leuschner, J., Schmidt, M., Baguer, D.O., Maass, P., 2021. Lodopab-ct, a benchmark dataset for low-dose computed tomography reconstruction. Scientific Data 8, 109.

Lewis, D.D., 1995. A sequential algorithm for training text classifiers: Corrigendum and additional data, in: Acm Sigir Forum, ACM New York, NY, USA. pp. 13–19.

Li, Y., Li, K., Zhang, C., Montoya, J., Chen, G.H., 2019. Learning to reconstruct computed tomography images directly from sinogram data under a variety of data acquisition conditions. IEEE transactions on medical imaging 38, 2469–2481.

Lu, S., Yang, B., Xiao, Y., Liu, S., Liu, M., Yin, L., Zheng, W., 2023. Iterative reconstruction of low-dose ct based on differential sparse. Biomedical Signal Processing and Control 79, 104204.

Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R., 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM 65, 99–106.

Pelt, D.M., Batenburg, K.J., Sethian, J.A., 2018. Improving tomographic reconstruction from limited data using mixed-scale dense convolutional neural networks. Journal of Imaging 4, 128.

Roy, N., McCallum, A., 2001. Toward optimal active learning through sampling estimation of error reduction. int. conf. on machine learning .

Sagara, Y., Hara, A.K., Pavlicek, W., Silva, A.C., Paden, R.G., Wu, Q., 2010. Abdominal ct: comparison of low-dose ct with adaptive statistical iterative reconstruction and routine-dose ct with filtered back projection in 53 patients. American Journal of Roentgenology 195, 713–719.

Sener, O., Savarese, S., 2017. Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489 .

Settles, B., Craven, M., Ray, S., 2007. Multiple-instance active learning. Advances in neural information processing systems 20.

Seung, H.S., Opper, M., Sompolinsky, H., 1992. Query by committee, in: Proceedings of the fifth annual workshop on Computational learning theory, pp. 287–294.

Shen, Z., Wang, Y., Wu, D., Yang, X., Dong, B., 2020. Learning to scan: A deep reinforcement learning approach for personalized scanning in ct imaging. arXiv preprint arXiv:2006.02420 .

Shepp, L.A., Logan, B.F., 1974. The fourier reconstruction of a head section. IEEE Transactions on Nuclear Science 21, 21–43. doi:10.1109/TNS.1974.6499235.

Smailagic, A., Costa, P., Noh, H.Y., Walawalkar, D., Khandelwal, K., Galdran, A., Mirshekari, M., Fagert, J., Xu, S., Zhang, P., et al., 2018. Medal: Accurate and robust deep active learning for medical image analysis, in: 2018 17th IEEE international conference on machine learning and applications (ICMLA), IEEE. pp. 481–488.

Song, B., Shen, L., Xing, L., 2023. Piner: Prior-informed implicit neural representation learning for test-time adaptation in sparse-view ct reconstruction, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1928–1938.

Strong, D., Chan, T., 2003. Edge-preserving and scale-dependent properties of total variation regularization. Inverse problems 19, S165.

Tancik, M., Mildenhall, B., Wang, T., Schmidt, D., Srinivasan, P.P., Barron, J.T., Ng, R., 2021. Learned initializations for optimizing coordinate-based neural representations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2846–2855.

Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R., 2020. Fourier features let networks learn high frequency functions in low dimensional domains. Advances in Neural Information Processing Systems 33, 7537–7547.

Tian, Z., Jia, X., Yuan, K., Pan, T., Jiang, S.B., 2011. Low-dose ct reconstruction via edge-preserving total variation regularization. Physics in Medicine & Biology 56, 5949.

Wang, C., Shang, K., Zhang, H., Zhao, S., Liang, D., Zhou, S.K., 2022. Active ct reconstruction with a learned sampling policy.

Willemink, M.J., Noël, P.B., 2019. The evolution of image reconstruction for ct—from filtered back projection to artificial intelligence. European radiology 29, 2185–2195.

Wu, D., Ren, H., Li, Q., 2020. Self-supervised dynamic ct perfusion image denoising with deep neural networks. IEEE Transactions on Radiation and Plasma Medical Sciences 5, 350–361.

Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z., 2017. Suggestive annotation: A deep active learning framework for biomedical image segmentation, in: Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20, Springer. pp. 399–407.

Yang, Q., Yan, P., Zhang, Y., Yu, H., Shi, Y., Mou, X., Kalra, M.K., Zhang, Y., Sun, L., Wang, G., 2018. Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss. IEEE transactions on medical imaging 37, 1348–1357.

Ye, D.H., Buzzard, G.T., Ruby, M., Bouman, C.A., 2018. Deep back projection for sparse-view ct reconstruction, in: 2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP), IEEE. pp. 1–5.

Zang, G., Idoughi, R., Li, R., Wonka, P., Heidrich, W., 2021. Intratomo: self-supervised learning-based tomography via sinogram synthesis and prediction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1960–1970.

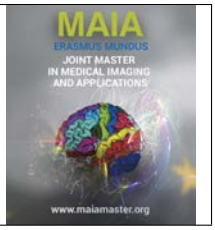
Appendix A. Hyperparameters

Table A.6: Hyperparameters for the base INR architecture.

Dataset	Loss function	Regularization weight	Learning rate	sigma	encoded size	MLP hidden layers
Phantom	MSELoss	0.0003	0.001	1	256	8
LoDoPaB-CT	Poisson	7	0.0005	4.5	256	8

Medical Imaging and Applications

Master Thesis, June 2023



Self-supervised pretraining for high-level feature extraction in computational pathology

Nohemí Sofía León Contreras, Marina D'Amato, Francesco Ciompi

Computational Pathology Group, Radboud University Medical Center, The Netherlands
sofialeoncontreras@gmail.com, marina.damato@radboudumc.nl, francesco.ciompi@radboudumc.nl

Abstract

The use of artificial intelligence, in particular deep learning, in computational pathology has enabled the development of computer-aided diagnostic tool systems which can help improve diagnostic accuracy and precision. However, obtaining high-quality pixel-level annotation in histopathology datasets can be very expensive and time-consuming. Through self-supervised learning meaningful image representation can be learned without the need for any annotations; downstream tasks such as classification can benefit from these learned representations. However, common end-to-end self-supervised learning approaches cannot be straightforwardly applied to whole-slide images due to their giga-pixel resolution. In this work, we explore a transformer-based architecture for the hierarchical aggregation of visual tokens called Hierarchical Image Pyramid Transformer using over 6,000 colorectal biopsy slides from the ExaMode project. HIPT uses two levels of DINO self-supervised learning to learn and aggregate rich image representation by taking advantage of the hierarchical structure of visual tokens across varying resolutions found in whole-slide images. Finally, a lightweight ViT is finetuned for a specified downstream tasks. We compare the performance of different pretraining regimes of HIPT on a binary and a multiclass classification task, where the best AUC ROC obtained are 0.956 and .897, respectively. For a qualitative evaluation of the embeddings learned by our self-supervised models, we provide UMAP scatter plots as a visual aid. These plots offer valuable insights into the separability and discriminative power of the learned features.

Keywords: Computational pathology, self-supervision, DINO, Vision Transformer, HIPT, Colorectal cancer

1. Introduction

The field of pathology is devoted to understanding the causes of disease (etiology) and the changes in cells, tissues, and organs that are associated with disease and give rise to the presenting signs and symptoms (pathogenesis) in patients. Pathology provides the scientific foundation for developing rational treatments and effective preventive measures through understanding the etiology and pathogenesis of the disease, thus providing the scientific foundation for the practice of medicine (Vinay Kumar, 2017). In clinical practice, pathology aims to obtain diagnostically important information in an objective and reproducible manner from microscopic images of tissue samples obtained via surgery, biopsy, or, less commonly, autopsy (Martín, 2021; Meijer et al., 1997). The tissue sample is thinly sliced with a microtome, mounted on a glass slide, and stained. Different cellular components can be revealed through different staining techniques; the most commonly used staining technique consists of hematoxylin and eosin (H&E). Hematoxylin is a basic dye that has affinity for acid (basophilic) structures of the cell, therefore it mainly stains the cell nuclei in a blue shade, while eosin is an acidic dye that binds to cytoplasmic (eosinophilic) structures of the cell staining them in multiple shades of pink (Martín, 2021; Slaoui et al., 2017). Another used technique is immunohistochemical (IHC) staining which allows for the detection of specific proteins expressed by the cells contained in the tissue section, therefore, facilitating the accurate identification of cells of specific origin (Ramos-Vara and Miller, 2014).

The information revealed through the examination of histopathology slides can help clinicians and re-

searchers to better understand the underlying pathology of a disease, improve the accuracy of diagnosis, and monitor the efficacy of treatments. Some techniques used include measuring morphological characteristics of cells and tissues, counting cell and tissue components, and using advanced methods such as cytometry and pattern recognition to identify subtle changes in tissue architecture or cell behavior (Meijer et al., 1997).

Computational pathology can be traced back to the 1960s, with the advent of computer technology, allowing for more reliable, easier measurements of cellular and tissue components and introducing more complex evaluations of their characteristics (van der Laak et al., 2021; Meijer et al., 1997). Nowadays, this term does not only encompass the high-resolution digitization of histopathology slides, with the first whole-slide scanners introduced to the field about two decades ago, but also the analysis of these images with computational tools for detection, segmentation, and diagnosis (Bera et al., 2019). The digitization of histopathology images has enabled the retrieval and analysis of information invisible to the human eye, which can help improve the accuracy and precision in grading systems and measurement in biomarker expression for personalized medicine (Laurinavicius et al., 2012). It also makes possible the automation of tissue-based diagnosis and quantification, thus potentially improving clinical workflows (Madabhushi and Lee, 2016). Due to the high resolution in which they are scanned, with .25 micrometers/pixel (40X) and .5micrometers/pixel (20X) being common digitization resolutions, whole-slide images (WSI) are large in size with one single slide of 20mm x 15mm scanned at 40X resulting in 80,000 x 60,000 pixels or 4.8Gp per channel (thus WSI being often referenced as gigapixel images). Research in digital pathology began employing traditional computer vision methods with handcrafted features like active contours, tissue texture features, nuclear shape, and size as described in Madabhushi and Lee (2016). More recent approaches have started using artificial intelligence (AI) such as machine-learning techniques, in particular Deep Learning (DL), as it allows the generation of robust algorithms that need fewer iterative optimizations for each dataset compared with methods where parameters are manually tuned (Abels et al., 2019). The construction of DL algorithms, rather than by explicit programming or by using predefined filters, yields powerful, hierarchical feature representations that, in most cases, outperform more traditional image analysis methods (van der Laak et al., 2021). The use of AI in this field is of special interest as whole-slide image scanning technology produces gigapixel-sized images (van der Laak et al., 2021), these algorithms can be used as a clinical decision support tool for precision diagnosis of the patient, easing the workload of clinicians by flagging suspicious regions or slides that may contain tumor cells for inspection, compute mitotic counts, improve accuracy and precision of IHC scoring or even apply standardized histological scoring (Abels et al., 2019). Additionally, the use of such models can reduce inter-observer variability (van der Laak et al., 2021).

One major obstacle that DL faces is that huge amounts of data are required; according to van der Laak et al. (2021) to address this problem, multicentric efforts have been carried out to increase the size of datasets and in that way cope with the variability in staining, scanning characteristics and tissue preparation across different laboratories. The use of supervised machine learning techniques, where learning entails the mapping of a set of input variables with their corresponding output variable to later apply such mapping to predict the output of unseen data (Cunningham et al., 2008), requires the use of ground truth labels. The ground truth labels may derive from patient outcome data (pathology report, laboratory information, or patient clinical history), from expert manual annotations of gigapixel whole-slide images (e.g. identifying cancer from benign tissue)(Abels et al., 2019).

Obtaining adequate, high-quality pixel-level annotated datasets can be very expensive and time-consuming, one strategy to alleviate this burden is to train algorithms in a weakly supervised manner where only slide-level labels are used (Lu et al., 2021). According to Zhou (2018), weak supervision can be grouped into three main categories:

- Incomplete Supervision: Only a small subset of the dataset is annotated, the rest remains unlabelled.
- Inexact Supervision: Only coarsed-grained labels are given, for example in histopathology images it would be desirable to have labels for every cell found in the image, but most of the time we have only WSI-level labels.
- Inaccurate Supervision: The labels available are not considered completely reliable (careless/weary annotator or images may be difficult to categorize).

In practice, these three categories are not mutually exclusive and they often occur simultaneously. One example of weak supervision is multiple-instance learning (MIL), where a DL framework operates on bags of embedding instances and label information is provided at the bag level but not at the instance level (Carbonneau et al., 2018). For example, in histopathological images, instead of using pixel-level annotations, patch-level or WSI-level annotations are provided (bag-level label). So the question arises of how to generate good instance representation without the need for any label. Through unsupervised learning, more specifically self-supervised pretraining, raw input data (images) can be used to generate meaningful learning signals without the need for a prior (Ciga et al., 2022). Through pretext tasks, self-supervised pretraining can provide richer

representations than when they are learned through supervised objectives (Caron et al., 2021). Commonly, after the self-supervised pretraining of the network with unlabelled data, the network is later finetuned on a downstream task. The better the self-supervision, the better the downstream performance (Newell and Deng, 2020). In recent years, there has been a significant acceleration in the progress of self-supervised pretraining, with methods being able to produce high-quality features that are comparable to or outperform those produced by ImageNet (Deng et al., 2009) pretraining. Features learned by ImageNet pretraining may transfer well to natural images. However, the domain shift to medical images pose concerns in relation to ImageNet pretraining, through self-supervised methods one can perform pretraining on the exact image distribution used for the downstream task, thus obtaining significant high-level feature representation of the data.

2. State of the art

2.1. Self-supervised learning methods

Different self-supervised methods for computer vision have proven their potential use with images with convolutional neural networks (CNN) and more recently with Vision Transformers. Below a review of some of these methods is presented.

2.1.1. Momentum Contrast (MoCo)

He et al. (2020) proposed a mechanism for building dynamic dictionaries for contrastive learning. The training is done by matching an encoded query to a dictionary of keys via a contrastive loss (Figure 1). The goal is for the encoded "query" to be similar to its matching "key" and dissimilar to others when performing a dictionary look-up. The dictionary is maintained as a queue of encoded data samples, where the current encoded representations of the batch are queued and the oldest are dequeued. As the dictionary of keys can be much larger than the mini-batch size and it is grown on the fly, it keeps encoded keys from preceding mini-batches. While the parameters in the query-encoder are updated by backpropagation, the queue set-up does not allow the key-encoder parameters to be updated by back propagation nor by just copying the parameters from the newly-updated query-encoder, otherwise the key representations that are in the queue belonging to previous minibatches would become inconsistent representations. They solve this issue by implementing a momentum-based moving average of the query encoder as the slowly progressing key encoder. Formally speaking the key encoder's parameters (θ_k) are updated by Eqn. 1, where $m \in [0, 1)$ is the momentum coefficient and θ_q are the query encoder parameters.

$$\theta_k \leftarrow m\theta_k + (1 - m)\theta_q \tag{1}$$

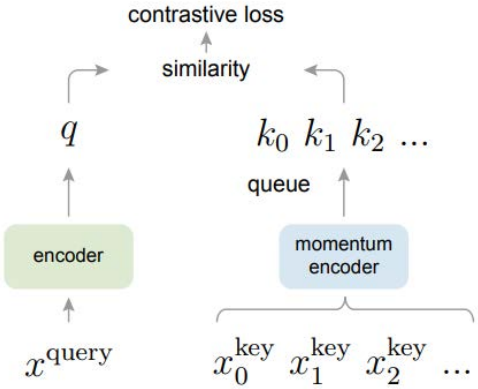


Figure 1: MoCo trains a visual representation encoder by matching an encoded query q to a dictionary of encoded keys using a contrastive loss (He et al., 2020).

This way, even though the old keys in the queue (from previous mini-batches) were encoded by previous versions of the key-encoder, the difference among the updates of the key-encoder can be made small. They prove that a large momentum ($m = 0.999$) works better than a smaller value, suggesting that a slowly evolving key encoder is relevant to making good use of a queue. After pretraining for 200 epochs a ResNet-50 encoder on ImageNet, they achieve 60.6% in top-1 classification accuracy using linear classification (a fully-connected layer followed by softmax) on frozen features.

2.1.2. Simple Framework for Contrastive Learning of Visual Representations (SimCLR)

The self-supervision framework proposed by Chen et al. (2020) learns representations by maximizing agreement between differently augmented views of the same data example via a contrastive loss in the latent space. It has four main components (see Figure 2): a data augmentation module that transforms any given data sample randomly to get two correlated views of the same example, a base neural network encoder that represents the augmented views, a small neural network (in their work they use a multi-layer perceptron with one hidden layer) that serves as a projection head, to project the representation into a smaller space, and a contrastive loss for the contrastive prediction task. The aim of the task is to identify the positive pair (the augmented views of the same image) given a set of data samples. The encoder is architecture agnostic, allowing for any choice of neural network architecture. Unlike MoCo, SimCLR does not train with a memory bank of representations instead it trains with large batch sizes with N ranging from 256 to 8192, since each sample of a batch yields 2 different augmented views (positive pair), it treats all other $2(N - 1)$ augmented samples as negatives examples, for example: a batch size of 8192, would give 16382 negative examples per positive pair. In their work, they prove that image augmentation op-

erator order is crucial for learning good representations, concluding that, for a contrastive task, composing a spatial/geometrical transformation with a color distortion transformation is critical to learn generalizable features, otherwise, the network might shortcut learn to differentiate different images based on their histogram alone. Classification accuracy is used as a proxy for representation quality. Similar to MoCo, they train a linear classifier on top of the frozen base encoder (ResNet-50) and achieve 76.5% on top-1 accuracy.

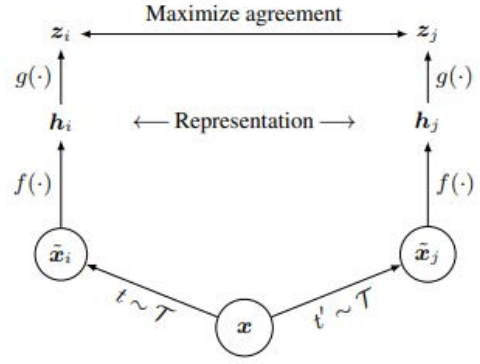


Figure 2: In SimCLR, two separate data augmentations are sampled from the same family of augmentations. The base encoder, $f(\cdot)$, and the projection head, $g(\cdot)$, are trained to maximize agreement. After training, only the encoder, $f(\cdot)$, is used for downstream tasks (Chen et al., 2020).

2.1.3. Bootstrap Your Own Latent (BYOL)

While MoCo and SimCLR rely on negative pairs, BYOL (Grill et al., 2020) does not need this, instead it bootstraps the outputs of a network to serve as targets for an enhanced representation. BYOL uses an online network and a target network, that interact and learn from each other. The online network is trained to predict the target's network representation of a different augmented view of the same image. BYOL follows a similar augmentation regime as SimCLR, with random crops with random horizontal flips (geometrical transformation), followed by a color distortion (random sequence of brightness, contrast, saturation, hue adjustments, and an optional grayscale conversion) with a final gaussian blur and solarization. Through ablation experiments, they show that BYOL is not as sensitive to the choice of data augmentation as SimCLR; due to BYOL's training task, the online network is incentivized to keep any information represented in the target network.

To avoid collapse they use a slow moving average of the online network's parameters as the target network's parameters, similar as the momentum encoder in MoCo but instead of using it to maintain consistent negative pair representations it is used to produce prediction targets for stabilizing the bootstrap step; this way, the target network represents a more stable version of the online network. The online network, is built by an encoder (any neural network architecture), a projector (a multilayer perceptron) and a predictor (a multilayer perceptron). The target network has the same encoder and projector architecture as the online network but no predictor. The goal is for the online network's predictor to match the target's projection of the different view of the same image as shown in Figure 3 by minimizing a similarity loss (MSE). Similar to SimCLR, after pretraining is finished, only the network's encoder is used for downstream tasks. The online network's encoder is used for downstream tasks. Using a ResNet-50 encoder, BYOL reaches 74.3% top-1 classification accuracy on ImageNet using a linear classification evaluation.

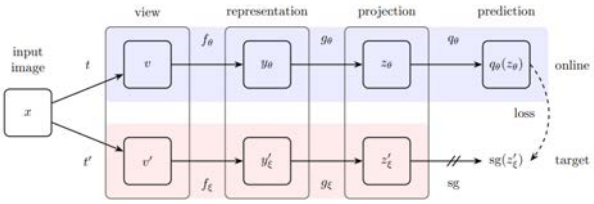


Figure 3: BYOL's goal is to learn a representation y_θ which can then be used for downstream tasks. Since the parameters in the target network (in red) are an exponential moving average of the online parameters, BYOL learns its representation by predicting previous versions of its outputs (Grill et al., 2020).

2.1.4. Self-distillation with No Labels (DINO)

Whereas the above-described self-supervised learning regimes are architecture agnostic and focused on using CNNs, Caron et al. (2021) explore how Vision Transformer (ViT) (Dosovitskiy et al., 2020) properties can be leveraged by self-supervised learning regimes since the success of Transformers in natural language processing is due to its use in self-supervised pretraining (Devlin et al., 2019). They propose a self-distillation with no labels self-supervised learning paradigm, that not only can exhibit explicit information about the semantic segmentation of an image when working with ViT but also works with CNNs by matching the state-of-the-art with a ResNet-50 encoder.

DINO draws inspiration from BYOL's framework, with the difference that both the online (in this case called student) and the target (called teacher in DINO) networks are the same architecture (encoder and projection head) and that cross entropy loss is used as the similarity loss function. DINO can be thought of as a Mean Teacher self-distillation, as both teacher and student have the same architecture, with no labels (Pham et al., 2022). Furthermore, since the teacher is being updated with an exponential moving average(ema) of the student, it can also be thought of as a codistillation, as two copies of the same model are trained in parallel (Anil et al., 2018).

From the same image sample, a set of augmented

views is built, where two of these views are global crops containing more than 50% of the image and the rest are local crops containing less than 50% of the image. Besides the global/local cropping, the images are augmented following the BYOL (Grill et al., 2020) augmentation regime. The student network sees all crops, while the teacher network only sees the two global crops, this regime encourages local-to-global correspondences.

When ViT is used as an encoder, an extra learnable token, called [CLS] token, is added to the token sequence (Dosovitskiy et al., 2020).The [CLS] token aggregates information from the entire token sequence, therefore its output is the one attached to the projection head. As observed in Figure 4, different random transformations of the same image sample x are passed through each of the networks. By normalizing the network's output with a softmax operation with temperature scaling τ, a k-dimensional output probability distribution $P(x)$ is obtained (Eqn.2).

$$P(x)^{(i)} = \frac{\exp(g_\theta(x)^{(i)}/\tau)}{\sum_k \exp(g_\theta(x)^{(k)}/\tau)} \quad (2)$$

Before the softmax operation, the output of the teacher network is centered, to avoid collapse induced by a dominant dimension, and sharpened, to avoid collapse induced by uniform distribution output. The centering of the teacher's output is done by adding a bias term to the teacher's outputs, said bias term is updated with an exponential moving average of the network's outputs, and sharpening is done by simply setting the teacher's temperature in the softmax operation to a low value compared to the student's temperature.

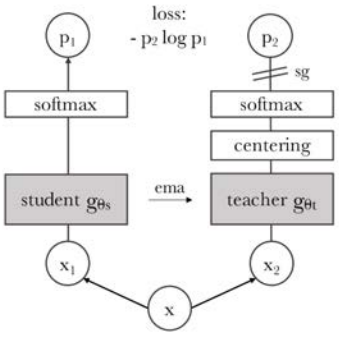


Figure 4: The teacher and student network (g_{θ_t} and g_{θ_s} respectively) are composed of a backbone encoder f (ViT or ResNet), and of a projection head h (a 3-layer multi-layer perceptron). The features used in downstream tasks are the backbone f features. (Caron et al., 2021).

The student network's parameters are updated by minimizing the cross entropy loss between these two distributions. The teacher network is updated in a similar fashion as in MoCo (Eqn. 1), with the difference that the momentum coefficient follows a cosine scheduling regime (in MoCo this value is a constant). With this parameter update regime, the teacher network constantly outperforms the student network, thus guiding the training. Given that the teacher network constantly outperforms the student, the encoder that is used for downstream tasks is the teacher's encoder.

Table 1: Top-1 accuracy on linear classification on the validation set of ImageNet. For the self-supervised methods, the linear classifier is learnt on frozen features of the encoder.

Method	Encoder Architecture	Top 1 (%)
Supervised	Resnet50	79.3
MoCo	Resnet50	60.6
SimCLR	Resnet50	76.5
BYOL	Resnet50	74.3
DINO	Resnet50	75.3
Supervised	ViT-S	79.8
BYOL	ViT-S	72.7
DINO	ViT-S	77.0

In linear classification evaluation, DINO surpasses the previous methods discussed obtaining a 75.3% top-1 accuracy when using a ResNet-50 encoder, as shown in Table 1, Caron et al. (2021) also compare the performance of ViT-small as the encoder and show it is almost on par with the supervised training accuracy. The Image-Net features obtained in DINO when using a ViT as the encoder can be used in a k-NN classifier and obtain a top-1 accuracy performance of 74.5% almost on par with a linear classifier (77% of top-1 accuracy), this property only emerges when using DINO with ViT as the encoder (Caron et al., 2021). Additionally, without the need for finetuning, the self-attention maps of the [CLS] token in the different heads of the last layer contain information about the semantic segmentation of an image.

2.2. Self-Supervised Learning in Digital Pathology

End-to-end approaches cannot straightforwardly be applied to WSIs, mainly because their gigapixel dimension makes them unable to fit into the memory of modern GPUs. Even switching to central processing unit (CPU) computation would not resolve this problem, as a single WSI can easily require tens of gigabytes of memory at full resolution (van der Laak et al., 2021). Another obstacle when working with WSIs is the low signal-to-noise ratio, where it is possible that only a small area of the image may contain relevant information for the image-level label. As discussed in the Introduction section, MIL can be used to address this problem, by detecting non-overlapping tissue patches that contain the true signal (e.g. cancerous cells) while suppressing the noisy ones (e.g healthy tissue) and then global pooling these instance-level representations to obtain a WSI-level embedding (Chen and

Krishnan, 2021; Ilse et al., 2018; Tellez et al., 2019). However, common MIL methods only take into account patterns found in individual patches of tissue, while ignoring the spatial relationships among them. To address this pitfall, Chen et al. (2022) proposed a new ViT-based architecture called Hierarchical Image Pyramid Transformer (HIPT), that exploits the natural hierarchical structure inherent in WSIs. HIPT consists of nested aggregations of visual tokens of incrementing resolution coming from two levels of DINO-pretrained ViTs and a weakly supervised ViT to learn high-resolution image representations, see Figure 5.

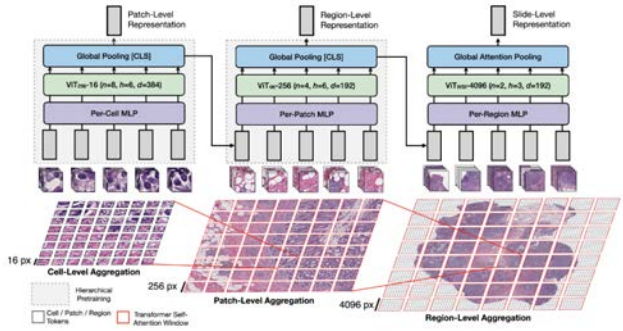


Figure 5: The nested aggregation of visual tokens in different resolution levels allow for representation of fine-grained morphological features as well as coarser-grained features like tissue-to-tissue relationships. (Chen et al., 2022).

HIPT exploits the fixed scale at a given magnification objective found in WSIs that natural images don't posses, therefore the visual tokens extracted using a ViT allow a consistent comparison of the visual elements represented. At magnification of 20X, the cell-level aggregation ViT, called $ViT_{256} - 16$, as in Dosovitskiy et al. (2020) works with 256x256 pixel tissue patches, modelling 16x16 visual tokens that can model different morphological features and individual cell. The patch-level aggregation ViT, called $ViT_{4096} - 256$, characterizes the interactions within the tissue microenvironment by using the [CLS] tokens given by $ViT_{256} - 16$ coming from the 256x256 patches contained in a region of 4096x4096. This way, Vision transformers using DINO pretraining regime are able to learn interpretable histological features, where the different attention heads learn distinct morphological phenotypes (Chen and Krishnan, 2021).

As of now, according to Chen and Krishnan (2021) there is a lack of diverse and well-curated pathology datasets that would enable generalization across diverse tissue and organ types in cancer pathology. In this work, we explore leveraging the work presented by Chen et al. (2022) where HIPT was pretrained using 10,678 FFPE H&E stained resection WSIs from The Genome Cancer Atlas (TCGA) dataset by further training the network on Radboudumc archival material by expanding the training domain to biopsy whole-slide images.

3. Material and methods

3.1. Datasets

The datasets used in this project are part of the ExaMode (Extreme-scale Analytics via Multimodal Ontology Discovery and Enhancement) project, which aims to provide automatic and semi-automatic methods to improve the efficiency and the effectiveness of the diagnoses in the pathology domain with the positive effect of reducing the pathologists' workload (Menotti L. and G., 2023). More specifically, the images used in this work for self-supervised pretraining are 8,868 H&E-stained colorectal biopsy slides cut from 6,563 paraffin blocks of 3,601 patients taken between the years 2000-2009 from Radboud University Medical Center (RUMC), Nijmegen, Netherlands. As the diagnoses reports refer to the paraffin blocks and not the individual slides, the slides from the same block were combined creating 'packed' slides.

Another set of 76 colorectal slides from Azienda Ospedaliera Cannizaro and Gravina Hospital Caltagirone ASP, Catania, Italy was used as part of the test set in a classification downstream task (slide-level weak supervision) to test the domain shift robustness of the features learned through self-supervised pretraining. This set had associated reports per slide. The reports were then labeled by student assistants into classes normal, hyperplastic polyps, low-grade dysplasia (LGD), high-grade dysplasia (HGD) and cancer.

Table 2: Data class distribution overview of colorectal biopses. Blocks (packed slides) from Nijmegen and individual slides from Catania. LGD:low grade dysplasia, HGD:high grade dysplasia

Medical Center	# Images	Normal	Hyperplastic	LGD	HGD	Cancer
RUMC	6563	3072	850	1505	204	303
Catania	76	13	1	35	16	11

3.1.1. Data Preprocessing

To remove excessive white space found in the WSI and make sure that pretraining of HIPT is primarily done in the tissue of the WSI, tissue segmentation and patch extraction were done following the methodology proposed by Lu et al. (2021). For the tissue segmentation, the WSI is converted from RGB to HSV color space, a median blur is applied to blur the edges, then a binary mask containing the tissue in the WSI is created by thresholding the saturation channel. A morphological closing operation to fill small holes is done to the resulting mask. The found contours are then filtered based on an area threshold. For patch extraction, the algorithm crops square patches of a specified size from within the segmented tissue at a specified magnification. The upper left corner coordinates of the extracted patches and the slide metadata are stored using the HDF5 hierarchical data format. For this work, it

was chosen to extract non-overlapping [4096 × 4096] regions at 20x(0.5 μm/pixel). Once the coordinates of these regions were obtained, the pixel regions were extracted from the WSI, and stored in HDF5 data format. Inside the HDF5 file, the regions were stored as a chunked dataset of 256 chunks each containing a [256×256] tissue patch to facilitate its access during pretraining of the $ViT_{256} - 16$, see Figure 6.a) for an overwiew of the preprocessing pipeline.

Table 3: Total number of 4096 × 4096 whole-slide image regions per class and per institution. LGD:low grade dysplasia, HGD:high grade dysplasia.

Institution	Class	Number of regions
Catania	Normal	236
	Hyperplastic	12
	LGD	931
	HGD	597
	Cancer	508
RUMC	Normal	175754
	Hyperplastic	29449
	LGD	82640
	HGD	23010
	Cancer	32987

After the preprocessing pipeline described above was run on the colon WSI datasets, a total of 346,124 regions were obtained, totaling $\sim 88.6M$ of [256 × 256] tissue patches. However, self-supervised pretraining was only done with the RUMC dataset, leaving 343,840 tissue regions for pretraining of the $ViT_{4096} - 256$ and $\sim 88M$ tissue patches for pretraining the $ViT_{256} - 16$. The Catania dataset was only used as a test set in slide-level weak supervision.

3.2. Proposed Method

The method followed in this work is the one proposed by Chen et al. (2022). It can be separated into two stages: self-supervised pretraining and slide-level weak supervision, both of these stages are explained below. Before going into the two stages of the method, it is essential to provide a formal overview of the Vision Transformer architecture since it is the backbone on which HIPT is built. Instead of using convolutional layers and expanding the field of view with the depth of the network as CNNs do, the ViT uses self-attention mechanisms and feed-forward neural networks to capture global relationships and dependencies within the visual input. Following Dosovitskiy et al. (2020), an image x with resolution $L \times L$ (or x_L where $x \in \mathbb{R}^{L \times L \times C}$, where C is the number of channels) is split into a sequence of non-overlapping flattened 2D patches of size $l \times l$, such that $x_l \in \mathbb{R}^{N \times (l^2 \times C)}$ where $N = L^2/l^2$ is the number of resulting patches. These patches are then embedded by a linear projection into a vector of size d. An extra learnable token, named [CLS] token, is prepended to the sequence of patch embeddings. Even though, the [CLS] token is not connected to any image patch, its state at the output of the ViT encoder can be used as the image representation therefore a classification head is attached to it when doing classification tasks. Learnable position embeddings are then added to the patch embeddings and [CLS] token to retain positional information. This sequence of embedded vectors is then processed by a sequence of cascading standard Transformer encoder blocks with multi-head self-attention as proposed by Vaswani et al. (2017). The self-attention layers update the token representations by looking at the other token representations with an attention mechanism.

This thesis will now proceed to discuss the two stages of self-supervised pretraining and the slide-level weak supervision stage, that form the HIPT architecture. When referring to the different ViT architectures used throughout the method, the following notation will be used: n for the number of transformer blocks in the ViT, h for the number of attention heads in each transformer block, and d for the vector dimension of the embeddings, this size stays constant through all of the ViT blocks and is the output vector dimesion of the [CLS] token of that ViT.

3.2.1. Self-supervised pretraning

As shown by Chen et al. (2022) and explained above in the State of the art section, HIPT architecture uses two levels of DINO-based knowledge distillation with ViT to learn high resolution hierarchical image representation. HIPT approaches WSI in a similar way as it is done in Natural Language Processing, where embeddings are aggregated at the character-, word-, sentence- and paragraph-level to create document representations.

The first level ViT, from here now referred to as $ViT_{256} - 16$, aggregates the information found in [16 × 16] non-overlapping pixel tokens found in an image patch of size [256×256], or x_{256}. At 20X magnification, a bounding box of x_{16} pixels is a bounding box of 8×8 micrometers which encompass visual concepts that are object-centric in featuring single cells, $ViT_{256} - 16$ aggregates these features to capture local clusters of cell-to-cell interactions. The $ViT_{256} - 16$ architecture used in this work has the next characteristics: n=12, h=6, d=384.

The DINO pretraining of $ViT_{256} - 16$ takes place in a similar fashion as proposed by Caron et al. (2021), a student and teacher network with identical architectures are built and different augmented views of sampled WSI patches are fed to them. Each WSI patch, x_{256}, is augmented into 10 different views, 2 global crops of size x_{224}, and 8 local crops of size x_{96}. The teacher network only sees the global crops whereas the student sees the ten crops. The following augmentations are randomly

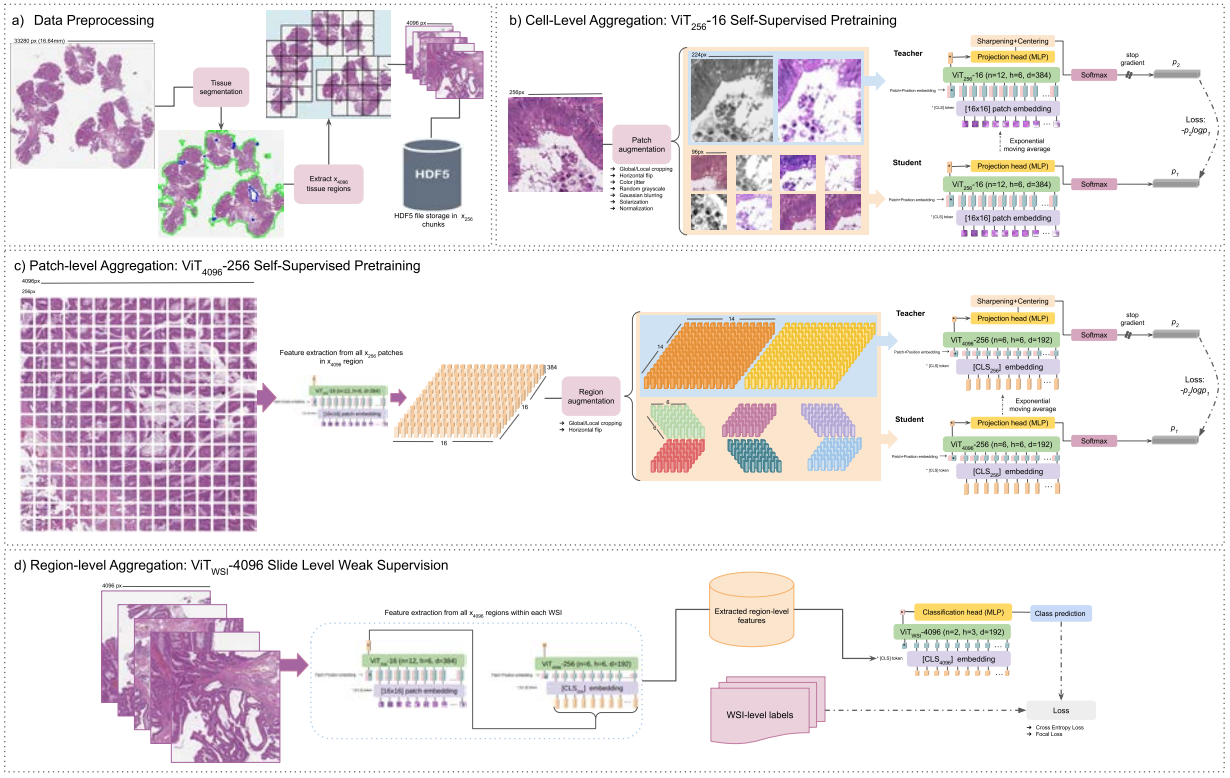


Figure 6: Overview of the proposed method showing the data preprocessing, the DINO-based self-supervised aggregation layers and the slide level weak supervision. By using aggregation layers to aggregate visual tokens at varying image resolutions both the tissue microenvironment as well as the interactions between clusters of cells can be modeled. x_L indicates the size of the squared WSI image patch.

applied to the crops: horizontal flip (p=0.5), color jitter (p=0.8), random grayscale (p=0.2), gaussian blurring with a standard deviation of 0.1, solarization with a threshold of 128 (p=0.2), and normalization. These augmentations help the model capture essential features by making the model robust to variations in object orientations, noisy images, lighting conditions, and color distributions, thus enabling the models to achieve higher levels of performance and reliability.

The $[CLS]_{256}$ (subindex denotes ViT input image resolution that is being aggregated at that level) token output vector of size d=384 of the $ViT_{256}-16$ is then passed through a projection head consisting of a 3-layer multilayer perceptron (MLP) with hidden dimension 2048 and an output dimension of 65,536. Since ViT does not use any batch normalization, the projection head is also built without any batch normalization. The output feature vector of the projection head is centered and sharpened only in the teacher network and then normalized with a temperature softmax over the feature dimension in both networks as depicted in Figure 6.b). The distribution similarity of the teacher and student feature vector is then measured with a cross entropy loss. Through a stop-gradient operator on the teacher, backpropagation to minimize this loss is only applied to update the parameters of the student model with AdamW optimizer. The teacher parameters are updated through an exponential moving average of the student's parameters.

The second level ViT, from here now referred to as $ViT_{4096}-256$, aggregates the feature information from non-overlapping x_{256} patches found in an image region of size $[4096\times4096]$, or x_{4096}, hence its name of patch-level aggregation ViT in Figure 6.c). $ViT_{4096}-256$ aggregates the cell-to-cell interactions represented in $[CLS]_{256}$ to characterize macro-scale interactions between clusters of cells and their organization in tissue. The extracted $[CLS]_{256}$ tokens pertaining to the 256 non-overlapping x_{256} patches within the x_{4096} region are reshaped into a $16\times16\times384$ 2D feature grid, this helps to retain the positional location of the x_{256} patches within the x_{4096} region. The $ViT_{4096}-256$ architecture used in this work has the next characteristics: n=6, h=6, d=192.

As for DINO pretraining data augmentation regime, 2 global crops are made with a size of $14\times14\times384$ and 8 local crops with a size of $6\times6\times384$, retaining the scale of the crops done for the pretraining of $ViT_{256}-16$. As additional data augmentation, horizontal flip (p=0.5) is applied to the crops. The DINO-based pretraining of $ViT_{4096}-256$ follows an almost identical approach as the pretraining of $ViT_{256}-16$ described above, only difference is that at the beginning of the $ViT_{4096}-256$ architecture the linear embedding layer with added position embeddings is done to the 256 $[CLS]_{256}$ token

feature vectors to produce a 256 set of 192-dim embeddings. This setup intuitively retains the input sequence length of tokens of the ViT as the image size scales, therefore making the computational complexity of aggregating big WSI regions the same as aggregating small WSI patches.

3.2.2. Slide-Level Weak Supervision

The last level ViT in HIPT, from here now referred to as $ViT_{WSI} - 4096$, aggregates the region-level representations obtained from $ViT_{4096} - 256$. $[CLS]_{4096}$ token aggregation of regions from the same WSI is done via formulating a slide-level classification task $P(y|WSI)$, where y is a slide-level label. $ViT_{WSI} - 4096$ has the following architecture: n=2, h=3, d=192. Due to not all x_{4096} regions obtained during preprocessing being continuous in the WSI, positional embeddings are ignored.

The whole hierarchical image pyramid transformer formulation can be written as:

$$\begin{aligned} HIPT(x_{WSI}) &= ViT_{WSI} - 4096\left(\left\{CLS_{4096}^{(k)}\right\}_{k=1}^{M}\right) \\ \rightarrow CLS_{4096}^{(k)} &= ViT_{4096} - 256\left(\left\{CLS_{256}^{(j)}\right\}_{j=1}^{256}\right) \\ \rightarrow CLS_{256}^{(j)} &= ViT_{256} - 16\left(\{x_{16}^{(i)}\}_{i=1}^{256}\right) \end{aligned} \tag{3}$$

where 256 is the sequence length of both non-overlapping [16x16] pixel tokens found in x_{256} images patches and non-overlapping [256x256] patches found in x_{4096}, and M is the total number of x_{4096} regions found in a given WSI. The average number of M in the dataset used in this work is 52 regions per WSI.

Two slide-level classification tasks were formulated to finetune $ViT_{WSI} - 4096$ on the previously extracted $[CLS]_{4096}$ tokens from the pretrained $ViT_{4096} - 256$; a binary classification problem and a 4-class classification problem. For the binary classification, Normal and Hyperplastic colorectal WSI were grouped into a 'Normal' class, whereas LGD, HGD and cancer colorectal WSI were grouped into an 'Abnormal' class. For the 4-class classification, Normal and Hyperplastic colorectal WSI were grouped into a 'Normal' class and the other three classes were LGD, HGD, and cancer.

The 'packed' slides from RUMC were split into train (5116), validation (1321), and test(72) sets. The WSI from Catania (76) were used exclusively for testing. In Table 4 the number of colorectal slides from each class and each center per split are shown.

Table 4: Train, validation, and test split done for slide-level weak supervision. LGD:low grade dysplasia, HGD:high grade dysplasia.

Medical Center	Split	Class	# WSI
RUMC	Training	Normal	2909
		Hyperplastic	651
		LGD	1188
		HGD	146
		Cancer	222
	Validation	Normal	743
		Hyperplastic	184
		LGD	291
		HGD	43
		Cancer	60
	Test	Normal	14
		Hyperplastic	12
		LGD	20
		HGD	12
		Cancer	14
Catania	Test	Normal	13
		Hyperplastic	1
		LGD	35
		HGD	16
		Cancer	11

3.3. Proposed Experiments

Experiment Notation: Several $ViT_{256} - 16$ and $ViT_{4096} - 256$ were pretrained for this work, the name of the datasets that any given ViT has been pretrained on, whether in this work or in the work from Chen et al. (2022) will be put before the ViT notation. That being the case, the $ViT_{256} - 16$ and $ViT_{4096} - 256$ pretrained in Chen et al. (2022) from now on will be referenced as $TCGA\ ViT_{256} - 16$ and $TCGA\ ViT_{4096} - 256$, respectively.

Pretraining $ViT_{256} - 16$ and $ViT_{4096} - 256$ at the same time is way too computationally expensive, therefore the pretraining is done in stages.

3.3.1. $ViT_{256} - 16$ Self-Supervised pretraining

In the work of Chen et al. (2022), $ViT_{256} - 16$ is pretrained with 104M tissue patches from TCGA across 33 cancer types and a batch size of 256 for 400,000 iterations (thus only pretraining for 1 epoch). Given that we have a similar number of colorectal tissue patches ($\sim 88M$) extracted from RUMC dataset, it was chosen to pretrain $ViT_{256} - 16$ in this work also for one epoch. A base learning rate of 0.0005 was used, with the first 10% of the epoch used to warm up to the base learning rate followed by decay using a cosine schedule.

In the context of this work, two different approaches for the DINO-based pretraining of $ViT_{256} - 16$ were taken, as depicted in Figure 7.a:

a. $TCGA+RUMC\ \ ViT_{256}-16$: Pretraining $ViT_{256}-16$ by finetuning from the available $TCGA\ ViT_{256}-16$ weights.

b. $RUMC\ ViT_{256} - 16$: Pretraining $ViT_{256} - 16$ from random initialization of the network's weights with the x_{256} patches from RUMC.

After pretraining of $ViT_{256}-16$, inference is run on the ViT encoder belonging to the teacher network with a batch size of 256 so all the x_{256} patches belonging to the same region x_{4096} are tokenized in the same forward pass. The resulting feature tensor of size $[256 \times 384]$ is saved as a .pt file for pretraining of the next level ViT.

$ViT_{256}-16$ is the only ViT in HIPT that will actually 'see' the image pixels, since the ViTs of the next levels work by aggregating the features extracted from the transformer of the previous stage.

3.3.2. $ViT_{4096}-256$ *Self-Supervised Pretraining*

In the work of Chen et al. (2022), $ViT_{4096}-256$ is pretrained with 408,218 tissue patches from TCGA and a batch size of 256 for 200,000 iterations, meaning that they trained for 125 epochs. This could be interpreted as the $ViT_{4096}-256$ having seen 51,200,000 'unique' regions. So calculating the number of epochs the $ViT_{4096}-256$ needs to be trained can be obtained by dividing 51,200,000 between the number of x_{4096} regions. We chose to follow this intuition for two reasons: 1) training by following the same number of iterations reported by Chen et al. (2022) makes the $ViT_{4096}-256$ training batch size dependent and 2) a batch size of 256 was not supported by the available GPU memory in our workstations. There were 343,825 regions extracted from RUMC's colorectal WSI dataset, which results in having to pretrain $ViT_{4096}-256$ for 149 epochs, for simplicity this number was rounded up to 150 epochs. Learning rate scheduling was the same as in $ViT_{256}-16$ pretraining, where the first 15 epochs were used to warm up to the base learning rate followed by decay using a cosine schedule.

It was chosen to work with the extracted $[CLS]_{256}$ tokens from the above-described $ViT_{256}-16$ pretraining schemes. Additionally, we also chose to extract the $[CLS]_{256}$ tokens from $TCGA\ ViT_{256}-16$ to pretrain $ViT_{4096}-256$, based on the underlying assumption that the single-cell morphological feature aggregation learned by $ViT_{256}-16$ can be thought to possess a certain level of generalizability across various tissue types, hence only pretraining the $ViT_{4096}-256$, which learns to aggregate clusters of cell-to-cell interactions, will still yield meaningful representations. Therefore, two different approaches per $ViT_{256}-16$ encoder (six different $ViT_{4096}-256$ pretraining experiments in total) to perform the DINO-based pretraining of $ViT_{4096}-256$ were done, as depicted in Figure 7.b:

- Using $[CLS]_{256}$ tokens from $TCGA\ ViT_{256}-16$:
  a. $RUMC\ ViT_{4096}-256$: Pretraining $ViT_{4096}-256$ from random initialization of the network's weights.
  b. $TCGA + RUMC\ ViT_{4096}-256$: Pretraining $ViT_{4096}-256$ by finetuning from the available $TCGA\ ViT_{4096}-256$ weights.

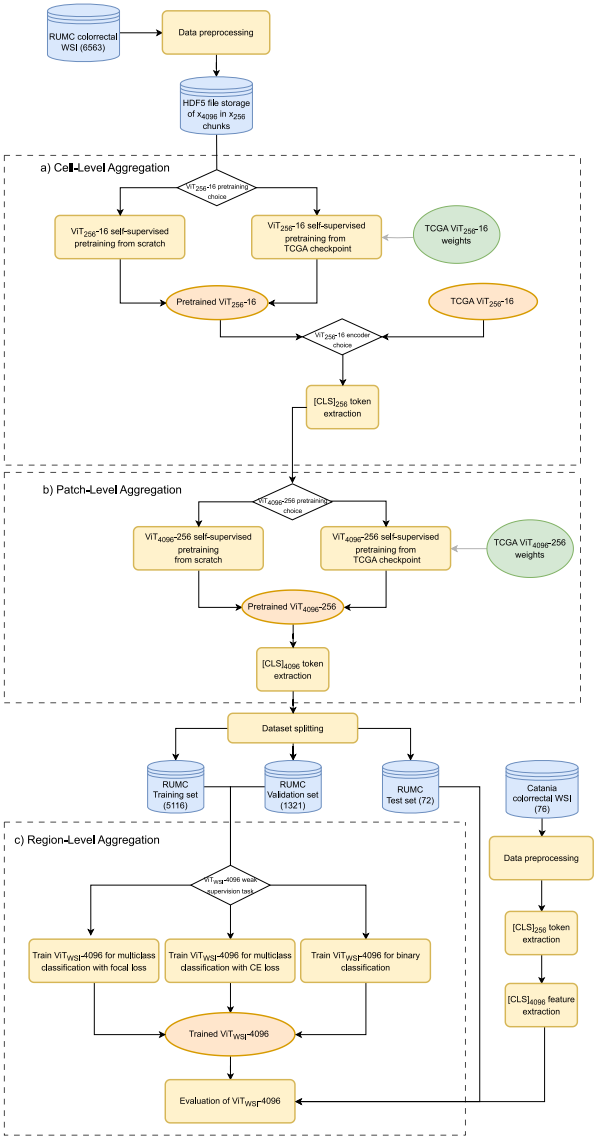


Figure 7: Proposed experiments flowchart. Each stage of HIPT is encased in dotted boxes.

- Using $[CLS]_{256}$ tokens from $TCGA + RUMC\ ViT_{256}-16$:
  a. $RUMC\ ViT_{4096}-256$: Pretraining $ViT_{4096}-256$ from random initialization of the network's weights.
  b. $TCGA + RUMC\ ViT_{4096}-256$: Pretraining $ViT_{4096}-256$ by finetuning from the available $TCGA\ ViT_{4096}-256$ weights.
- Using $[CLS]_{256}$ tokens from $RUMC\ ViT_{256}-16$:
  a. $RUMC\ ViT_{4096}-256$ Pretraining $ViT_{4096}-256$ from random initialization of the network's weights.
  b. $TCGA + RUMC\ ViT_{4096}-256$ Pretraining $ViT_{4096}-256$ by finetuning from the available $TCGA\ ViT_{4096}-256$ weights.

After pretraining of $ViT_{4096}-256$ is done, inference is run on the ViT encoder belonging to the teacher network for each x_{4096} region. All the 192-dim $[CLS]_{4096}$ tokens belonging to regions of the same WSI are then saved in a .pt file for their use in the slide-level weak supervision.

3.3.3. $ViT_{WSI}-4096$ Weak Supervision

In order to have an evaluation baseline, induction was run on $TCGA\ ViT_{256}-16$ and $TCGA\ ViT_{4096}-256$ to obtain the $[CLS]_{4096}$ token representation of the train, validation, and test sets. $ViT_{WSI}-4096$ was trained for 50 epochs using the Adam optimizer with a batch size of 1 with 32 gradient accumulation steps, and a fixed learning rate of 0.0002.

For the binary classification cross entropy loss was used, and for multiclass classification two experiments were run one using cross entropy loss and another one using focal loss. This same setup of training one $ViT_{WSI}-4096$ for each classification task (see Figure 7.c) was used for the 6 different combinations of $ViT_{256}-16 + ViT_{4096}-256$ explained above.

3.3.4. Distributed Training

In order to handle the amount of data in this work, every pretraining experiment was done across multiple GPUs at the same time in a process called distributed data-parallel (DDP) training. This training paradigm allows for the efficient utilization of computational resources and enables faster training on large datasets. To use DDP training in PyTorch, the model needs to be wrapped with PyTorch's DDP module. This ensures that the model's parameters are correctly synchronized during training across the distributed replicas.

During DDP process initialization, the model is replicated across multiple GPUs and parameter gradients are organized into buckets to improve communication efficiency. During the forward pass each replica processes locally a different subset of the training data in a similar manner as it would happen in single-machine,single-GPU training. Loss is calculated locally in each of the processes. During the backward pass, the gradients are computed independently of the DDP process, DDP uses autograd hooks registered at construction time to trigger gradient synchronization: once all the gradients in the same bucket are ready, a collective communication operation, called *all_reduce*, computes the average of said gradients across all instances. This way the gradient field of each corresponding parameter across all DDP processes is the same when the optimizer updates the weights in each local model.

3.4. Feature evaluation

To qualitatively assess the features learned in the pretraining of HIPT, we chose to use Uniform Manifold Approximation and Projection (UMAP), a non-linear dimensionality reduction technique renowned for visualizing high-dimensional features. By generating UMAP scatter plots, we can effectively reduce the dimensionality of the learned features while preserving their underlying intrinsic structure, allowing us to visualize the separability and discriminative power of the pretrained encoders. It was chosen to plot the $[CLS]_{256}$ tokens obtained from the three different $ViT_{256}-16$ used in this work, as several tissue types can be found at WSI and x_{4096} region level, and furthermore, it might be harder to give labels to the feature abstraction done by $ViT_{4096}-256$.

Since the RUMC and Catania datasets only contain WSI-level labels, patch-level features were extracted from CRC-100K (Kather et al., 2018), a publicly available patch-level labeled dataset with 100,000 non-overlapping x_224 image patches at 20X magnification with H&E staining of human colorectal cancer (CRC) and normal tissue. The tissue patches found in CRC-100K are annotated with the following 9 non-overlapping tissue classes: adipose, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma, and colorectal adenocarcinoma epithelium.

Following the work of Caron et al. (2021), where they show that self-supervised ViTs can learn semantic segmentation of the scene layout, we extract and visualize independently the multi-head self-attention from the two self-supervised ViTs to show the fine-grained visual concepts such as cell locations and coarse-grained visual concepts such as broader tumor cellularity learned by $ViT_{256}-16$ and $ViT_{4096}-256$ respectively. Additionally, factorized attention distributions combining $ViT_{256}-16$ and $ViT_{4096}-256$ attention distributions are also given. We compare these attention maps to the tissue segmentations obtained by following the method proposed by Bokhorst et al. (2023).

4. Results

As stated above, self-supervised pretraining of $ViT_{256}-16$ and $ViT_{4096}-256$ was done with distributed data-parallel training in PyTorch using NCCL backend for distributed GPU training.

$TCGA + RUMC\ ViT_{256}-16$ was trained using 3 NVIDIA A100-40GB with a batch size per GPU of 128. $RUMC\ ViT_{256}-16$ was trained on 2 GeForce RTX 2080 Ti with a batch size per GPU of 32. Each of the 6 different proposed $ViT_{4096}-256$ pretraining schemes were trained on 2 GeForce RTX 2080 Ti with a batch size per GPU of 32.

4.1. Slide-Level Weak Supervision

In this subsection the results obtained in the different classification tasks will be shown. The training of all the different 21 classification experiments was done using only one GPU with the hyperparameters explained in

section 3.3.3, each classification experiment took two hours of wall time to complete.

The different experiments will be referred to with the names of the different combinations of $ViT_{256}-16$ and $ViT_{4096}-256$ encoders. For example, the evaluation baseline that uses the $TCGA\ ViT_{256}-16$ and $TCGA\ ViT_{4096}-256$ for cell-level aggregation and patch-level aggregation respectively will be called $TCGA\ ViT_{256}-16, TCGA\ ViT_{4096}-256$.

4.1.1. Binary Classification

For the binary classification task, it was chosen to measure accuracy, precision, recall, F1 score, and AUC ROC (Area Under the Curve of the Receiver Operating Characteristics). The classification results obtained by $TCGA\ ViT_{256}-16, TCGA\ ViT_{4096}-256$ experiment baseline yielded an accuracy of 0.84 with an AUC ROC of 0.93.

Table 5: Binary classification results. AUC ROC: Area Under the Curve of the Receiver Operating Characteristics

Experiments	Accuracy	Precision	Recall	F1 score	AUC ROC
TCGA $ViT_{256}-16$, TCGA $ViT_{4096}-256$	0.84459	**0.98851**	0.79630	0.88205	0.93194
TCGA $ViT_{256}-16$, TCGA+RUMC $ViT_{4096}-256$	0.87162	0.97849	0.84259	0.90547	**0.95602**
TCGA $ViT_{256}-16$, RUMC $ViT_{4096}-256$	0.65541	0.96721	0.54630	0.69822	0.92315
TCGA+RUMC $ViT_{256}-16$, TCGA+RUMC $ViT_{4096}-256$	0.83108	0.96629	0.79630	0.87310	0.95394
TCGA+RUMC $ViT_{256}-16$, RUMC $ViT_{4096}-256$	**0.90541**	0.94340	**0.92593**	**0.93458**	0.94444
RUMC $ViT_{256}-16$, TCGA+RUMC $ViT_{4096}-256$	0.72297	0.92405	0.67593	0.78075	0.91042
RUMC $ViT_{256}-16$, RUMC $ViT_{4096}-256$	0.61486	0.94737	0.50000	0.65455	0.84190

When using $TCGA\ ViT_{256}-16, TCGA+RUMC\ ViT_{4096}-256$ an improvement in all metrics but precision (due to the precision/recall trade-off) can be observed, being this experiment the one with the highest AUC ROC (see Figure 8 where all ROC curves are plotted). Further improvement in accuracy, recall and F1 score is seen in experiment $TCGA+RUMC\ ViT_{256}-16, RUMC\ ViT_{4096}-256$ while still achieving a higher AUC ROC than the baseline experiment (0.944). A deterioration in the performance of the binary classifier can be appreciated when using the $RUMC\ ViT_{256}-16$ encoder with any combination of the pretrained $ViT_{4096}-256$ as it can be noted in Figure 8 and in Table 5. $RUMC\ ViT_{256}-16, RUMC\ ViT_{4096}-256$ being the experiment with the worst performance across all metrics. Another experiment where the performance of the binary classification is low is $TCGA\ ViT_{256}-16, RUMC\ ViT_{4096}-256$ where all obtained metrics are lower than the baseline experiment.

4.1.2. Multiclass Classification

For the multiclass classification task, it was chosen to measure top-1 accuracy, top-2 accuracy, balanced accuracy, quadratic Cohen's Kappa, F1 score, and AUC ROC. For this task, it was chosen to train $ViT_{WSI}-4096$ with cross entropy loss and with focal loss. The results of each experiment are shown in Table 6, the results obtained with each loss function will be commented on separately, comparison between their performances will be made in the Discussion Section.

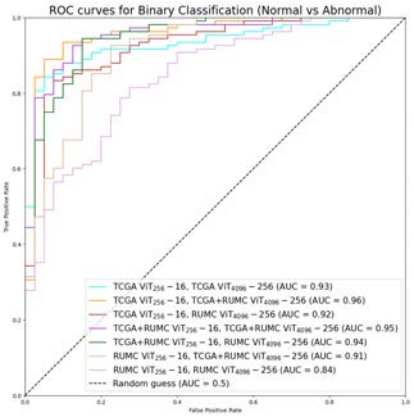


Figure 8: Receiver Operating Characteristics curves of the seven binary classification experiments.

Cross Entropy Loss. The classification results obtained by $TCGA\ ViT_{256}-16, TCGA\ ViT_{4096}-256$ experiment baseline yielded a Cohen's Kappa of 0.68 with a macroaveraged AUC ROC of 0.86. Figure 9 presents the ROC curves obtained using the One-vs-Rest (OVR) approach for multiclass classification, to obtain the One-vs-Rest ROC of a given class, said class is treated as the positive class while considering the rest of the classes as the negative class, once all the OVR ROC curves are computed their individual AUC ROC is averaged and macro-averaged AUC ROC is obtained.

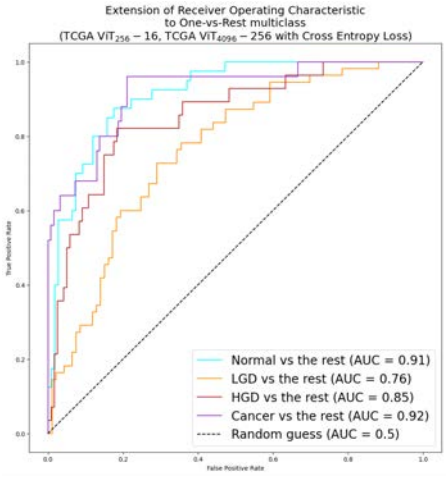


Figure 9: Performance of the baseline multiclass classification model trained with cross entropy loss in distinguishing each single class from the remaining classes. LGD:low grade dysplasia, HGD:high grade dysplasia.

This baseline classifier holds the highest top-1 accuracy, balanced class accuracy, Cohen's Kappa, and F1 score obtained across all experiments performed using cross entropy loss. When using $TCGA+RUMC\ ViT_{256}-16, TCGA+RUMC\ ViT_{4096}-$

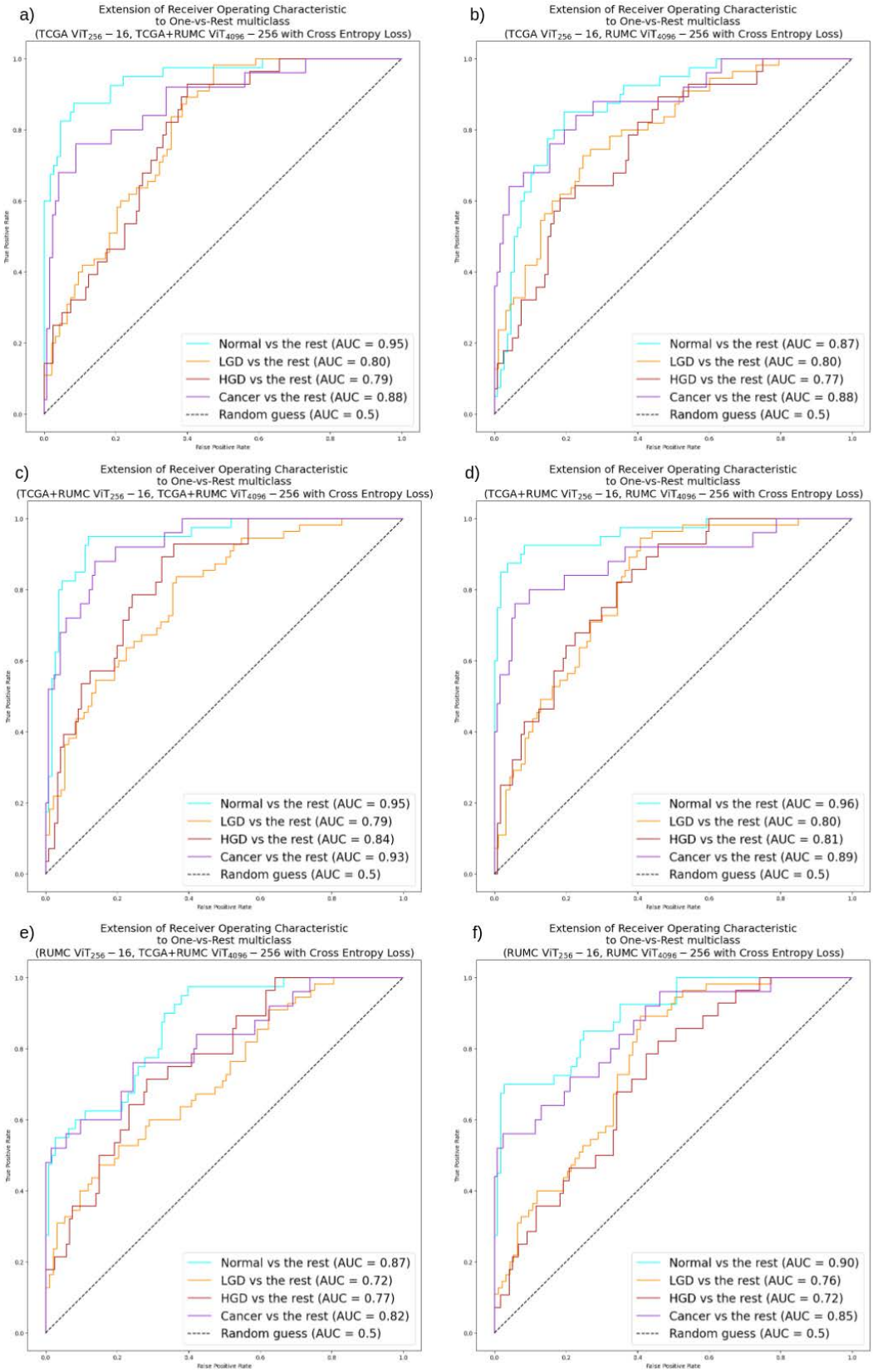


Figure 10: Performance of the different multiclass classification models trained with cross entropy loss in distinguishing each single class from the remaining classes. Each row of plots represents a different $ViT_{256} - 16$ encoder. LGD:low grade dysplasia, HGD:high grade dysplasia.

Table 6: Multiclass Classification results. F1 score was calculated per class and then the weighted mean was found. AUC ROC reported is macro AUC ROC, where AUC ROC is computed independently for each class and all class-wise AUC ROC curves are averaged. AUC ROC: Area Under the Curve of the Receiver Operating Characteristics

	Cross Entropy Loss						Focal Loss					
Experiment	Top-1 Accuracy	Top-2 Accuracy	Balanced Accuracy	Cohen's Kappa	F1 score	ROC AUC	Top-1 Accuracy	Top-2 Accuracy	Balanced Accuracy	Cohen's Kappa	F1 score	AUC ROC
TCGA $ViT_{256}-16$, TCGA $ViT_{4096}-256$	**0.63514**	0.86486	**0.62091**	**0.68444**	**0.63023**	0.86294	**0.69595**	**0.93243**	**0.68739**	0.71120	**0.69381**	0.87488
TCGA $ViT_{256}-16$, TCGA+RUMC $ViT_{4096}-256$	0.62838	0.81757	0.58943	0.68432	0.60937	0.85437	0.63514	0.89189	0.59680	0.71594	0.63868	0.85557
TCGA $ViT_{256}-16$, RUMC $ViT_{4096}-256$	0.54054	0.83108	0.52937	0.47907	0.52074	0.82875	0.52703	0.83784	0.53547	0.51661	0.51252	0.82717
TCGA+RUMC $ViT_{256}-16$, TCGA+RUMC $ViT_{4096}-256$	0.57432	**0.87162**	0.56266	0.56921	0.56147	**0.87721**	0.70270	0.89865	0.66636	**0.73202**	0.67887	**0.89733**
TCGA+RUMC $ViT_{256}-16$, RUMC $ViT_{4096}-256$	0.62162	0.81757	0.58284	0.61056	0.60018	0.86572	0.67568	0.81081	0.61823	0.62896	0.64328	0.85897
RUMC $ViT_{256}-16$, TCGA+RUMC $ViT_{4096}-256$	0.52703	0.72297	0.51278	0.45840	0.50604	0.79295	0.50676	0.72297	0.49510	0.39781	0.48241	0.78382
RUMC $ViT_{256}-16$, RUMC $ViT_{4096}-256$	0.62162	0.81081	0.56643	0.62941	0.61431	0.80694	0.52027	0.69595	0.56594	0.49599	0.50312	0.80065

256 results in an increase in the macro averaged AUC ROC and in top-2 accuracy. Experiments $TCGA+RUMC\ ViT_{256}-16, RUMC\ ViT_{4096}-256$ and $TCGA\ ViT_{256}-16, TCGA+RUMC\ ViT_{4096}-256$ are able to achieve a similar AUC ROC performance as the baseline experiment.

Same as in the binary classification experiments, the model performance decreases in experiments using $RUMC\ ViT_{256}-16$ and in experiment $TCGA\ ViT_{256}-16, RUMC\ ViT_{4096}-256$. This decrease in performance can also be appreciated in Figure 10.b), Figure 10.e), and Figure 10.f), where the ability of the classifiers to distinguish between normal vs the rest of the classes is lower than the other three trained classifiers depicted in the same Figure 10.

Focal Loss. The classification results obtained by $TCGA\ ViT_{256}-16, TCGA\ ViT_{4096}-256$ experiment baseline yielded a Cohen's Kappa of 0.71 with a macro averaged AUC ROC of 0.87. Figure 11 presents the ROC curves obtained using the One-vs-Rest (OVR) approach for multiclass classification, high AUC ROC for Normal vs the rest and Cancer vs the rest can be observed. Similar to the multiclass classifiers trained with cross entropy loss, this baseline classifier holds most of the highest performance metrics (top-1 accuracy, top-2 accuracy, balanced class accuracy and F1 score) among all other focal loss experiments.

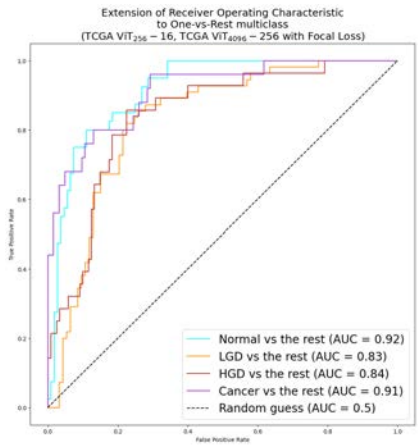


Figure 11: Performance of the baseline multiclass classification model trained with focal loss in distinguishing each single class from the remaining classes. LGD:low grade dysplasia, HGD:high grade dysplasia.

$TCGA\ ViT_{256}-16, TCGA+RUMC\ ViT_{4096}-256$ achieves a similar Cohen's Kappa and AUC ROC than the baseline experiment. $TCGA+RUMC\ ViT_{256}-16, RUMC\ ViT_{4096}-256$ also obtains a similar macro-averaged AUC ROC. This similarity in AUC ROC in $TCGA\ ViT_{256}-16, TCGA+RUMC\ ViT_{4096}-256$ and $TCGA+RUMC\ ViT_{256}-16, RUMC\ ViT_{4096}-256$ may be due to the high Normal vs the rest ROC curve exhibited in Figure 12.a) and Figure 12.d), respectively.

When using $TCGA+RUMC\ ViT_{256}-16, TCGA+RUMC\ ViT_{4096}-256$ for the classification task results in an increase in the macro averaged AUC ROC and in Cohen's Kappa accuracy. Observing the OVR ROC curves of the baseline experiment in Figure 11 and of $TCGA+RUMC\ ViT_{256}-16, TCGA+RUMC\ ViT_{4096}-256$ in Figure 12.c) we can notice an increase in the ability of the classifier to distinguish WSI labeled as normal tissue, WSI labeled as containing low grade dysplasia and WSI labeled as cancer with respect to the baseline experiment.

Once again experiments using $RUMC\ ViT_{256}-16$ display the worst perfomance with Cohen' Kappa below 0.5 and macro-averaged AUC ROC of 0.78 and 0.8 in $RUMC\ ViT_{256}-16, TCGA+RUMC\ ViT_{4096}-256$ and $RUMC\ ViT_{256}-16, RUMC\ ViT_{4096}-256$ respectively.

4.2. Feature Evaluation

As explained in the Material and methods section, we chose to use UMAP scatter plots to qualitatively evaluate the features learned by $ViT_{256}-16$. We chose to extract the features of the 100,000 labeled WSI patches of the CRC-100K image dataset, since the datasets that were used for this work only contain WSI-level labels. The UMAP-scatter plots of each of the 3 different $ViT_{256}-16$ encoders used for the presented experiments are shown in Figure 13. Figure 13.a) shows the UMAP of the features extracted from $TCGA\ ViT_{256}-16$, each different class is in a different color, it can be observed an overlap in the feature projection of cancer-associated stroma and smooth muscle. In Figure 13.a) the UMAP of the features extracted from $TCGA+RUMC\ ViT_{256}-16$ displays a more wide distribution of the class clusters in the UMAP space. Finally, in Figure 13.c) the UMAP scatter plot of the features from $RUMC\ ViT_{256}-16$ encoder show the biggest overlap between class clusters.

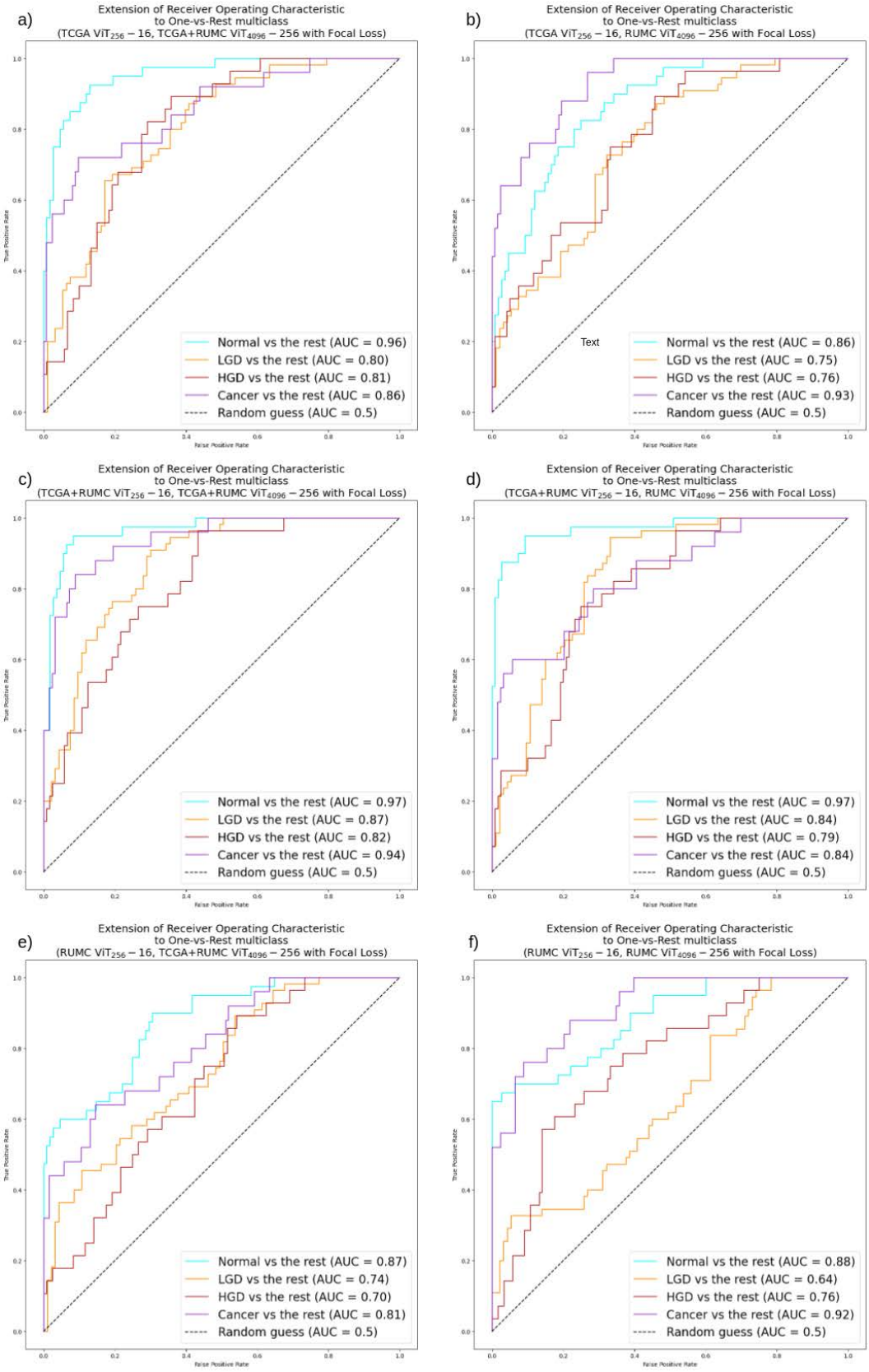


Figure 12: Performance of the different multiclass classification models trained with focal loss in distinguishing each single class from the remaining classes. Each row of plots represents a different $ViT_{256} - 16$ encoder. LGD:low grade dysplasia, HGD:high grade dysplasia.

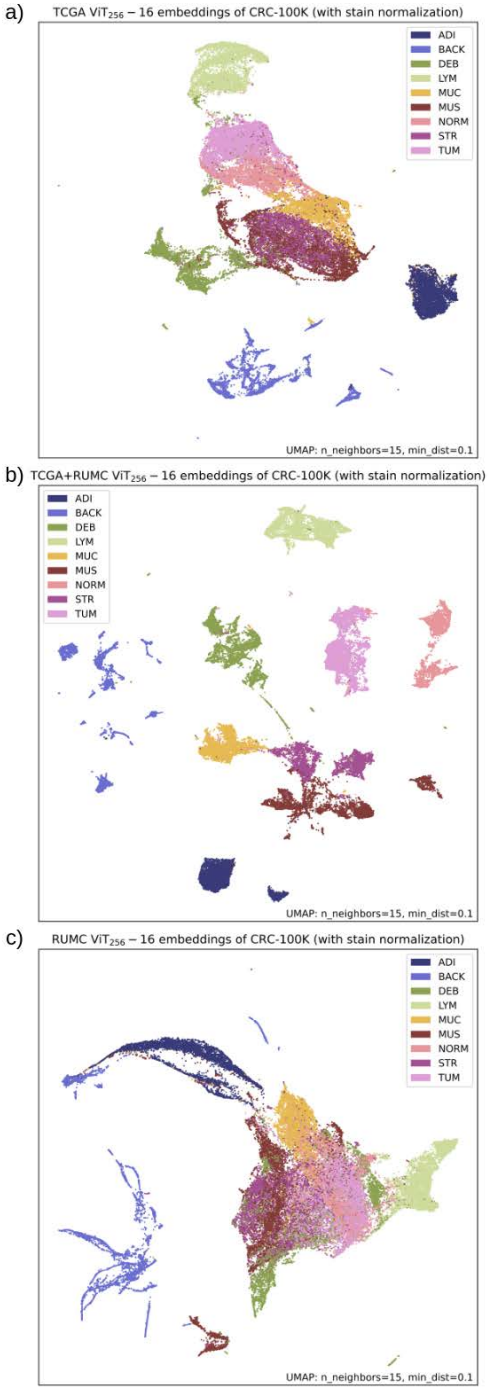


Figure 13: UMAP Visualization of pretrained embeddings from a) $TCGA\ ViT_{256}$ – 16, b)$TCGA$ + $RUMC\ ViT_{256}$ – 16 and c)$RUMC\ ViT_{256}$ – 16. ADI:adipose tissue, BACK:background, DEB:debris (Deb), LYM: lymphocytes, MUC: mucus, MUS: smooth muscle, NORM: normal colon mucosa (Norm), STR: cancer-associated stroma, TUM: colorectal adenocarcinoma epithelium.

An image region from the RUMC containing tumor was selected to showcase the multi-head self-attention visualization of the self-supervised ViT encoders. Since the $TCGA+RUMC\ ViT_{256}-16$, $TCGA+RUMC\ ViT_{4096}-256$ was the experiment with the highest AUC ROC it was chosen to use it for the attention visualization. The main tumor delineation obtained in the segmentation mask in Figure 14.b can be appreciated in the factorized self-attention of $TCGA+RUMC\ ViT_{256}-16$ and $TCGA+RUMC\ ViT_{4096}-256$ shown in Figure 14.f)

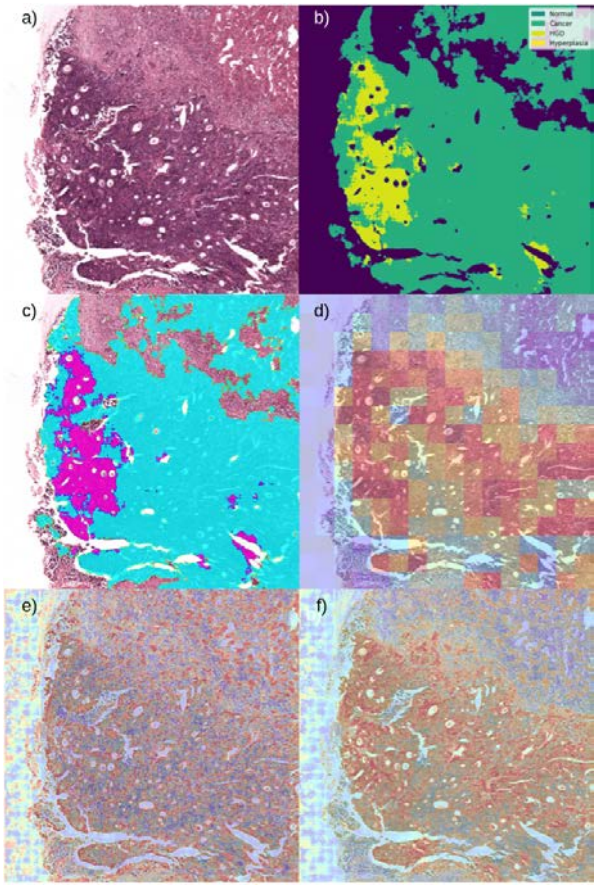


Figure 14: Hierarchical Attention Maps for colorectal cancer. a)Image region with colorectal cancer, b) segmentation mask obtained from Bokhorst et al. (2023) method, c) Overlay of segmentation mask on tissue region, d) Multi-Head Self-Attention of $ViT_{4096}-256$, e) Multi-Head Self-Attention of $ViT_{256}-16$, f) Factorized self-attention distributions of $ViT_{256}-16$ and $ViT_{4096}-256$.

5. Discussion

Performance improvement in the different HIPT models trained in this work happens when using the cell-level aggregation ViT model that was finetuned from the TCGA checkpoint with the RUMC data. This may point to considering these weights that have been pretrained using 10,678 H&E-stained diagnostic slides across 33 cancer types from TCGA as a good starting point for the development of cancer-specific encoders.

Before discussing the results of the different classification experiments, a bit more context on the RUMC dataset might be useful for understanding them. The RUMC dataset consists of 'packed' biopsy slide files

that contain several WSI cut from the same paraffin block, in some of these cases the relevant information for the slide label may be present in only one of the slides that were packed, therefore making the signal-to-noise ratio very low, this might explain the low recall performance, compared to the high precision one. This low signal-to-noise ratio can also explain the poor performance of the HIPT models trained solely on the RUMC data. Whereas the TCGA dataset is an extensively well-curated dataset of tissue resections where the WSI are composed of at least 80% tumor nuclei, hence having a higher signal-to-noise ratio and therefore aiding in the correct, discriminative abstraction of the aggregated token.

After further examining the results obtained in the binary classification task, an overall performance improvement can be noted when finetuning $ViT_{256} - 16$ using RUMC data. Even if an experiment that uses the baseline $ViT_{256}16$ encoder, $TCGA\ ViT_{256} - 16, TCGA + RUMC\ ViT_{4096} - 256$, is the one with the highest AUC ROC; $TCGA + RUMC\ ViT_{256} - 16, TCGA + RUMC\ ViT_{4096} - 256$ still has a similar AUC ROC performance too and $TCGA + RUMC\ ViT_{256} - 16, RUMC\ ViT_{4096} - 256$ has the highest F1 score, recall and accuracy. All binary experiments performed exhibit a high precision, meaning that even if they miss many positive cases thus yielding a low recall (as is the case in the baseline experiment and in $TCGA\ ViT_{256} - 16, RUMC\ ViT_{4096} - 256$), whenever a case is marked as Abnormal it is likely to be a true positive. It can be interpreted as the binary models being skeptic models: it is unlikely they flag a case to be abnormal hence missing to flag some abnormal cases, but when they do, it must have been because of overwhelming evidence that the case is in fact an abnormal case.

The OVR ROC curves, in multiclass classification, provide insights into the model's ability to correctly identify a given class while minimizing false positive predictions across the rest of the classes. In all multiclass experiments, Normal-vs-the rest and Cancer-vs-the rest AUC ROC are high meaning that the models are good at distinguishing these classes from the rest. The models tend to get confused between LGD and HGD which are the classes with more samples in the test set. Two different loss functions were used to train the $ViT_{WSI} - 4096$, cross entropy and focal loss. Through the use of a focusing parameter, the focal loss function focuses more on the predictions that the model is not very confident in while down-weighting the loss value for well-classified or very confident correct predictions, hence ensuring that predictions on hard examples improve over time rather than the model becoming overly confident with the easy ones. In most of the experiments (the ones using $RUMC\ ViT_{256} - 16$ excluded) we notice a significant improvement in the multiclass metrics, therefore, confirming the focal loss efficiency to address class imbalance during training.

Seeing that across all classification tasks, the performance of the models using $RUMC\ ViT_{256} - 16$ dramatically decreases, we conclude that it is crucial to have a good $ViT_{256} - 16$ encoder to have a good performance when doing WSI-classification.

The ability of the self-supervised pretrained $ViT_{256} - 16$ to successfully capture relevant patterns in the data can be appreciated in the UMAP scatter plots. In visualizing UMAP scatter plots of pre-extracted $ViT_{256} - 16$ features, the baseline $TCGAViT_{256} - 16$ already exhibits a high ability in capturing discriminative features per class. Clusters per class can be seen even if more than half of the class are very close to each other in the represented space, having smooth muscle and cancer-associated stroma completely overlapping. This could be interpreted as these classes having a poor patch-level representation or that these two classes share similarities. Considering the UMAP scatter plot of the features encoded by $TCGA + RUMC\ ViT_{256} - 16$, where these two classes are well differentiated into two separate clusters but still retain a high level of proximity to one another, we are inclined to assume that a combination of both factors may be the case. Moreover, in the UMAP scatter plot of the features encoded by $TCGA + RUMC\ ViT_{256} - 16$, we can see less scattering of each cluster class and higher distribution of the clusters across the UMAP visualization, which suggests that the self-supervised learning regime of $TCGA + RUMC\ ViT_{256} - 16$ has captured specific aspects of the diversity distribution found in colon data. Following this thread of thought, intuition might tell us that the $RUMC\ ViT_{256} - 16$ encoder will then have learned better representations of different tissue types encountered in the colon and rectum as it has been pretrained with only colon data. However, a poor class cluster differentiation can be noted when looking at the UMAP scatter plot of the embeddings obtained from the same images. It is possible that there is not enough tissue type diversity (or it is way too unbalanced) in the RUMC dataset to make the $RUMC\ ViT_{256} - 16$ learn meaningful distinctive features. These observations further support the notion that using the pretrained TCGA weights as a starting point for the hierarchical pretraining of HIPT is a sensitive way to approach the creation of cancer-specific encoders, due to $TCGA\ ViT_{256} - 16$ and $TCGA\ ViT_{4096} - 256$ possessing a certain level of generalizability across various tissue types.

Lastly, through factorized attention visualization of the output of the $TCGA + RUMC\ ViT_{256} - 16, TCGA + RUMC\ ViT_{4096} - 256$ experiment we can observe high attention areas (in red and orange) that seem to be attending to high tumor cellularity regions as the obtained segmentation mask shows. However further inspection of these attention maps has to be done by trained pathologists to give a correct assessment of the accuracy and quality of the features being attended by the ViTs in the

hierarchical pretraining of HIPT.

6. Conclusions

In this work, we explored the use of self-supervised knowledge in ViTs to obtain hierarchical representations of gigapixel images through the use of Hierarchical Image Pyramid Transformer architecture. The importance of variety in the data used for self-supervised pretraining has proven to be crucial for its good performance in different downstream tasks. It seems to be possible to obtain semantic segmentation from the factorized self-attention distributions. The presented work reaffirms the potential of transformer-based architectures to model complex, hierarchical relationships in data which can be extended into other domains of medical imaging/data.

6.1. Future work

One of the main challenges faced in this project was handling the huge amounts of data we had, which ascended to more than 10TB. More efficient ways to store data or the use of cloud services should be explored. Another way to tackle the burden of the huge amounts of data needed for pretraining and in order to mitigate the low signal-to-noise ratio in datasets where there is no pixel-level or patch-level labels could be done by sampling a subset of the x_{256} patches that represents the variety of tissue contained in the whole dataset. This could be done by extracting the features of the patches with a lightweight pretrained encoder, applying k-Means to these features, and then sampling a percentage of patches from each centroid. This data sampling approach would keep the pretraining method completely self-supervised where only the number of clusters to be made and the percentage to sample from each centroid would be given.

The recent release of DINOv2 could be implemented into the self-supervised hierarchical pretraining of HIPT architecture. Further ablation studies of the HIPT architecture could be done by varying the size of the region image from a range of 1024 to 4096.

Acknowledgments

I would like to thank my supervisors, Francesco and Marina, for all the time, attention and encouragement during my master's thesis project, for granting me the opportunity to work with them, and for helping me out when I needed it. I would like also to extend my gratitude to the MAIA program for granting me this amazing opportunity to study this master's degree, live in different countries and meet amazing people. To all my friends, whether in Mexico or Europe, for always being a source of inspiration, a helping hand, and a place to laugh. A mis padres, mi hermana y mis abuelos que sin ellos y sin su infinito apoyo y amor no estaría donde estoy hoy.

References

Abels, E., Pantanowitz, L., Aeffner, F., Zarella, M.D., van der Laak, J., Bui, M.M., Vemuri, V.N., Parwani, A.V., Gibbs, J., Agosto-Arroyo, E., et al., 2019. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the digital pathology association. The Journal of Pathology 249, 286–294.

Anil, R., Pereyra, G., Passos, A., Ormandi, R., Dahl, G.E., Hinton, G.E., 2018. Large scale distributed neural network training through online distillation. arXiv preprint arXiv:1804.03235 .

Bera, K., Schalper, K.A., Rimm, D.L., Velcheti, V., Madabhushi, A., 2019. Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology. Nature Review Clinical Oncology 16, 703–715. doi:10.1038/s41571-019-0252-y.

Bokhorst, J.M., Nagtegaal, I.D., Fraggetta, F., Vatrano, S., Mesker, W., Vieth, M., van der Laak, J., Ciompi, F., 2023. Deep learning for multi-class semantic segmentation enables colorectal cancer detection and classification in digital pathology images. Scientific Reports 13, 8398.

Carbonneau, M.A., Cheplygina, V., Granger, E., Gagnon, G., 2018. Multiple instance learning: A survey of problem characteristics and applications. Pattern Recognition 77, 329–353.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers, in: Proceedings of the International Conference on Computer Vision (ICCV).

Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F., 2022. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16144–16155.

Chen, R.J., Krishnan, R.G., 2021. Self-supervised vision transformers learn visual concepts in histopathology. Learning Meaningful Representations of Life, NeurIPS 2021 .

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. icml. arXiv preprint arXiv:2002.05709 .

Ciga, O., Xu, T., Martel, A.L., 2022. Self supervised contrastive learning for digital histopathology. Machine Learning with Applications 7, 100198.

Cunningham, P., Cord, M., Delany, S.J., 2008. Supervised learning. Machine learning techniques for multimedia: case studies on organization and retrieval , 21–49.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 .

Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al., 2020. Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems 33, 21271--21284.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning, in: Proceedings of

the IEEE/CVF conference on computer vision and pattern recognition, pp. 9729--9738.

Ilse, M., Tomczak, J., Welling, M., 2018. Attention-based deep multiple instance learning, in: International conference on machine learning, PMLR. pp. 2127--2136.

Kather, J.N., Halama, N., Marx, A., 2018. 100,000 histological images of human colorectal cancer and healthy tissue. URL: https://doi.org/10.5281/zenodo.1214456, doi:10.5281/zenodo.1214456.

van der Laak, J., Litjens, G., Ciompi, F., 2021. Deep learning in histopathology: the path to the clinic. Nature Medicine 27, 775--784. doi:10.1038/s41591-021-01343-4.

Laurinavicius, A., Laurinaviciene, A., Dasevicius, D., Elie, N., Plancoulaine, B., Bor, C., Herlin, P., 2012. Digital image analysis in pathology: benefits and obligation. Analytical cellular pathology 35, 75--78.

Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F., 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering 5, 555--570.

Madabhushi, A., Lee, G., 2016. Image analysis and machine learning in digital pathology: Challenges and opportunities. Medical image analysis 33, 170--175.

Martín, D.T., 2021. Advancing Computational Pathology with Deep Learning: From Patches to Gigapixel Image-Level Classification. Phd thesis. Radboud University. Nijmegen, NL.

Meijer, G., Beliën, J., Van Diest, P., Baak, J., 1997. Origins of... image analysis in clinical pathology. Journal of clinical pathology 50, 365.

Menotti L., M.S., G., S., 2023. The examode ontology,v2.0. doi:/10.5281/zenodo.7669237.

Newell, A., Deng, J., 2020. How useful is self-supervised pretraining for visual tasks?, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7345--7354.

Pham, M., Cho, M., Joshi, A., Hegde, C., 2022. Revisiting self-distillation. arXiv:2206.08491.

Ramos-Vara, J., Miller, M., 2014. When tissue antigens and antibodies get along: revisiting the technical aspects of immunohistochemistry|the red, brown, and blue technique. Veterinary pathology 51, 42--87.

Slaoui, M., Bauchet, A.L., Fiette, L., 2017. Tissue sampling and processing for histopathology evaluation. Drug Safety Evaluation: Methods and Protocols , 101--114.

Tellez, D., Litjens, G., van der Laak, J., Ciompi, F., 2019. Neural image compression for gigapixel histopathology image analysis. IEEE transactions on pattern analysis and machine intelligence 43, 567--578.

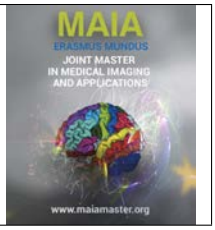
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. arXiv:1706.03762.

Vinay Kumar, Abul K. Abbas, J.C.A., 2017. Robbins Basic Pathology. Elsevier - Health Sciences Division, Philadelphia, PA.

Zhou, Z.H., 2018. A brief introduction to weakly supervised learning. National science review 5, 44--53.

Medical Imaging and Applications

Master Thesis, June 2023



Age Prediction from 3D Structural MRI Images

Stela Lila, Xavier Lladó and Arnau Oliver

VICOROB Research Institute, Girona, Spain

Abstract

The use of structural magnetic resonance imaging (MRI) data has allowed for an in-depth exploration of how the brain experiences age-related neuroanatomical changes. These transformations occur both locally and network-wide as it undergoes maturation and aging. Based on these alterations, minimizing the difference between the real and the chronological age, so-called brain age delta, necessitates accomplishing precise brain age, which can serve as a biomarker. In recent times, there has been a significant amount of researches conducted in the field of age prediction utilizing data from brain MRI scans. A multitude of studies have explored utilizing machine learning algorithms like support vector machines (SVMs) random forests (RFs), and deep learning approaches employing both 2D and 3D imaging modalities. In this thesis, we explore age prediction models based on familiar structural networks using convolutional neural networks (CNN) with volumetric data from 1,016 healthy subjects aged 50-98 years (ImaGenoma dataset). The model design incorporates preprocessing techniques to standardize the images, including bias correction, registration, brain extraction, and intensity normalization, ensuring consistent input for subsequent analysis. We tested the impact of different architectures on the ImaGenom dataset. Fine-tuning only the prediction block of SFCN architecture described in Peng et al. (2021a) achieved the best MAE of 3.33 years and r2 coefficient = 0.6713. Overall, more refined results and an increase in prediction metric was obtained when fine-tuning the hyperparameters of the networks.

Keywords: Brain MRI, age prediction, healthy subjects, machine learning, convolutional neural networks

1. Introduction

Studies using structural magnetic resonance imaging (MRI) have shown that the brain experiences significant age-related neuroanatomical changes over the span of the life process. One approach utilized by researchers to determine an individual's brain age involves quantifying gray matter volume over time. This methodology recognizes that as humans age in years they experience a gradual decrease in gray matter content due to diverse factors including both natural aging processes (Good et al., 2001), (Taki et al., 2011) and external environmental influences (Baxi et al., 2020) or neurodegenerative conditions (Thompson et al., 2003), (Fisher et al., 2008).

The field of age prediction has passed through many key milestones which significantly emphasize its potential applications in various domains. (Franke et al., 2009) introduced an innovative concept known as "brain age". Their methodology involved utilizing a machine learning algorithm to examine structural brain MRI data towards estimating one's chronological age. (Cole et al., 2010) introduced an innovative approach to predict brain age. They combined neuroimaging and machine learning methods in their study and discovered that differences in how our brains age can be associated with changes in cognitive abilities and the development of neurodegenerative conditions. In a study involving a significant number of healthy individuals, Koutsouleris et al. (2012) examined the accuracy of predicting brain age and its connection to schizophrenia. The results of their investigation indicated that estimating brain age might have the potential to be used as a biomarker for neurodevelopmental disorders like schizophrenia. 3 years later, in an exciting breakthrough, Liem et al. (2015) explored the world of functional connectivity patterns derived from resting-state functional MRI to predict brain age. Through their study, they discov-

ered a remarkable connection between functional connectivity and the passage of time, shedding new light on how our brains age and the underlying neural mechanisms involved. Kaufmann et al. (2017) undertook a comprehensive investigation by integrating multiple types of neuroimaging data to predict brain age in a diverse group of individuals. Their study emphasized the promising prospects of combining various imaging techniques, including structural MRI, diffusion MRI, and functional MRI. This multimodal approach demonstrated significant potential in enhancing the accuracy and dependability of brain age prediction models.

Recently, deep learning approaches have gained prominence in brain age prediction research. In a seminal study, Cole et al. (2019) introduced the BrainAGE framework, which employed deep neural networks to estimate brain age based on structural MRI data. Their research marked a significant milestone, as the BrainAGE framework surpassed previous methods in terms of performance and showcased the immense potential of deep learning techniques in accurately predicting brain age. Nowadays there are a lot of publications that made a major impact on the field of brain age prediction, worth noting are (Couvy-Duchesne et al., 2020), (Peng et al., 2021b), (Barbano et al., 2022) up to the most recent one published by Zhang et al. (2023). Some of the aforementioned papers will be exploited in section 2.

Given the necessity to develop a reliable approach for accurately predicting brain age and potentially assisting in the clinical scenario, the main aim of this research is to develop a robust method for brain age prediction. This can serve as a baseline for comparing an individual's biological age with their chronological age, uncovering disparities that may indicate accelerated aging or age-related diseases. Our work focused on 2 different deep-learning architectures. The first architecture tried was the one described by Yin et al. (2023) where the last output layer was changed to 2 neurons indicating the number of binary classification, while the second architecture was the SFCN model described by Peng et al. (2021b).

2. State of the art

Due to its clinical relevance, MRI quantification of the human brain has been widely investigated. Within the scope of this research, the methods implemented for age prediction of the brain can differ in different factors. These factors include the input data type such as 2D projections, 3D projections, 3D volumes, and 3D maps of gray and/or white matter. Another factor is the dataset used for each of the architectures. All of the methods based on the training process can be grouped into 2 main categories, machine learning, and deep learning approaches, all of which have state-of-the-art representations that can be seen in Table 1.

2.1. Machine learning approaches

Machine learning has become a remarkable tool that empowers computers to learn from data and make predictions or decisions without relying on explicit programming. At its core, machine learning revolves around the idea of prediction, wherein algorithms learn from past data to uncover valuable patterns, relationships, and trends. These insights can then be applied to anticipate and forecast future outcomes. During the training phase, the machine learning algorithm learns from a labeled dataset, where the input features (also known as predictors or independent variables) are associated with known target values (also known as labels or dependent variables) (Mitchell et al., 2007). Once the model is trained, it can be used for inference, where new, unseen data is fed into the model, and the algorithm makes predictions or decisions based on the learned patterns.

Various publications used different types of input data reflecting in a different MAE for each method. Commonly, as stated in Da Costa et al. (2020) and Baecker et al. (2021), authors used 3D volume and/or Voxel-based morphometry data as an input to their models. On the other hand, Jönemo et al. (2022) utilized 2D projections of 3D MRI volumes. The paper by Da Costa et al. (2020) emphasizes the use of shallow machine learning models and feature engineering techniques to achieve competitive performance. The work published by Baecker et al. (2021) compares the effectiveness of region-based and voxel-based morphometric data in machine learning models, highlighting the impact of feature representation choices. Jönemo et al. (2022) introduced an efficient approach by leveraging 2D projections of 3D MRI volumes, utilizing 2D CNNs for feature extraction, and reducing computational complexity. These differences showcase the diverse avenues explored by researchers to improve the accuracy, interpretability, and efficiency of brain age prediction models.

2.2. Deep learning approaches

Recently, deep learning techniques have made significant strides showcasing unparalleled levels of prediction accuracy. This innovation is able to outperform humans in certain scenarios offering valuable support for healthcare providers who are looking to make critical clinical decisions, (LeCun et al., 2015), (Cole et al., 2017), (De Fauw et al., 2018). In the context of prediction, deep learning models excel at capturing complex patterns and relationships, enabling them to make accurate predictions based on input features. The training process involves optimizing the model's weights and biases to minimize the prediction error and improve performance (LeCun et al., 2015). One challenge that produced a great impact on the field of deep learning was Predictive Analysis Challenge (PAC 2019). The

challenge specifically focused on utilizing T1-weighted brain MRIs to predict the age of subjects in multicenter datasets. This challenge included 2 parts: 1) to achieve the most accurate age prediction, as measured by mean absolute error (MAE), and (2) to achieve the best MAE while keeping the Spearman correlation r-value between the prediction error (brain age delta) and the actual age below 0.1. Two reviewed papers that took the 1st and 3d place respectively on this challenge are (Peng et al., 2021b) and (Couvy-Duchesne et al., 2020). They achieved an MAE of 2.9 years and 3.33 years whereas the paper from Da Costa et al. (2020) reached an MAE of 3.57 years. This paper used machine learning algorithms while the other twwo used deep learning architecture.

A notable change in the way loss function is calculated in most architectures is presented in (Barbano et al., 2022). The authors addressed the challenge of biases arising from multi-site datasets utilizing leveraging contrastive learning techniques. By incorporating a contrastive loss function, the model learns informative representations that mitigate site-specific biases, resulting in improved generalization capabilities for robust brain age prediction. On the other hand, the paper published by Zhang et al. (2023) stands out as the newest contribution in the field of age prediction. This paper focuses on tackling age-level bias, a critical concern in brain age prediction.

3. Material and methods

Since SFCN described in Peng et al. (2021b) provided good results in predicting age from MRI images, an adaptation of this model is used in the current work for the purpose of prediction based on 3D volume data. We also implemented a binary classification model for our task as a start-up work (Yin et al., 2023).

3.1. Dataset description

3.1.1. ImaGenoma

The dataset used in this work is ImaGenoma dataset provided by Hospital Universitari de Girona Doctor Josep Trueta located in Girona, Spain. This dataset consists of T1-weighted and T2-weighted images of adults aged 50-98 years. It is important to mention that the individuals participating in this study do not exhibit any signs of cognitive impairment or conditions that affect the functioning or structure of their brains. Initially, the dataset contained 1022 images and after a quality vision check, we ended up working with 1016 subjects where 610, 204, and 202 were used for training, validation, and testing respectively. In Figure 1, we can observe a graphical representation showcasing the distribution of ages. It can be seen that most of the cases are centered between the interval of 60 to 70 years old. A few cases are found in the extremes of the distribution. Together with the images, an excel file containing the id of the patients, diet score, age, and sex was provided.

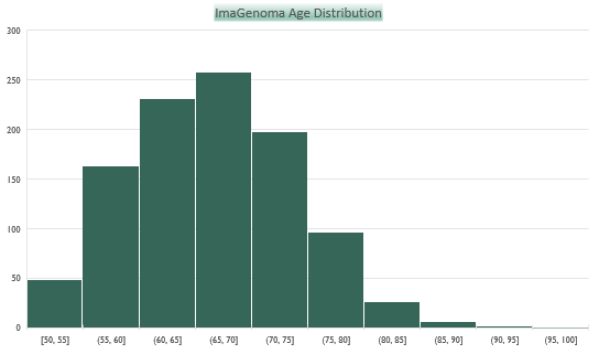


Figure 1: ImaGenoma age distribution.

Given that the images were not preprocessed, the following steps were undertaken to standardize the dataset: (i) applying bias field correction to all images, (ii) performing non-linear registration to MNI atlas, (iii) skull-stripping, and (iv) tissue and sub-cortical structures segmentation. More details about the preprocessing steps can be found in section 3.2.

3.1.2. UK Biobank

One of the deep learning approaches we followed was trained and tested using two different datasets: UK Biobank and PAC 2019. The main difference between them is in age distribution and number of subjects as can be seen in Table 2 and visualized in Figure 2. The brain imaging data obtained from the UK Biobank comprises a collection of multimodal brain scans primarily derived from a predominantly healthy cohort (Miller et al., 2016). The preprocessing pipeline for the UK Biobank data is explained in Alfaro-Almagro et al. (2018). The data release of the UK Biobank includes preprocessed data, eliminating the need for researchers to repeat the preprocessing pipeline. Access to the dataset is available to the researchers who present a project. To manage GPU memory requirements, the inputs are provided in 1mm MNI space and cropped to the central 160x192x160 voxels. Models are trained, validated, and tested separately using these processed inputs.

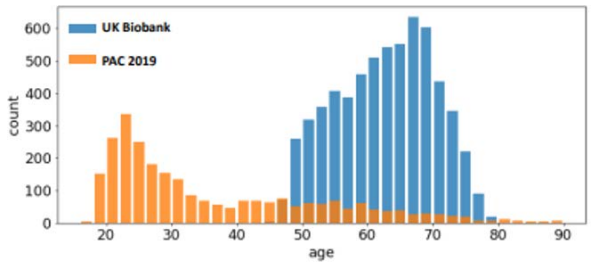


Figure 2: Age distribution of different datasets. *The UK Biobank (blue bars) and the PAC 2019 (orange bars)*

Table 1: Overview of the state-of-the-art methods and their respective Mean Absolute Error (MAE) used for brain age prediction.
SVR stands for support vector regression, DTR stands for decision tree regression, VBM stands for voxel-based morphometry, GM stands for gray matter and SFCN stands for simple fully convolutional network.

Methodology	Input	Author	Dataset	MAE (years)
Ensemble: SVR+DTR	3D Volume	**Machine Learning** Da Costa et al. (2020)	PAC 2019	4.57
SVR	VBM	**Machine Learning** Baecker et al. (2021)	UK Biobank	3.69
2D CNN	2D Projections	**Machine Learning** Jönemo et al. (2022)	UK Biobank	4.38
Inception V1	Priori info: GM	**Deep Learning** Couvy-Duchesne et al. (2020)	PAC 2019	3.33
SFCN	3D Volume	**Deep Learning** Peng et al. (2021b)	UK Biobank PAC 2019	**2.14** **2.9**
3D Resnet-18	3D Volume	**Deep Learning** Barbano et al. (2022)	OpenBHB	3.76
3D Resnet-34	3D Volume	**Deep Learning** Zhang et al. (2023)	UK Biobank OASIS ABIDE	2.55

Table 2: Differences between UK Biobank and PAC 2019 datasets.

Dataset	Age Range (yrs)	Age (yrs) Mean+STD	Number of Training/Validation/Test	Subject Total	Number of Sites
UK Biobank	42 – 82	52.7 + 7.5	12949/518/1036	14503	2
PAC 2019	17 – 90	35.9 + 16.2	2199/439/660	2638	17

3.1.3. PAC 2019

The dataset includes both a label-known training/validation dataset with a total of 2,638 subjects and a 'true' test set comprising 660 subjects. The labels of the test set are intentionally unknown from the competition participants, adding an element of uncertainty and evaluation to the competition. The subjects in the dataset are derived from 17 different sites. The majority of the data is based on the work by Cole and Franke (2017), with the organizers incorporating additional data from a few new sites. The input images are cropped to retain the central 160x192x160 voxels, used within UK Biobank data.

3.2. Preprocessing

Preprocessing plays a vital role in improving the quality and dependability of trained models. It encompasses a range of techniques, including data normalization, scaling, and feature extraction, all aimed at preparing raw data for optimal performance of the models. A key objective of preprocessing is to address inter-subject variability, ensuring that the data is uniform. This normalization enables fair and accurate comparisons across different individuals or samples. By reducing inter-subject variability through preprocessing, deep learning models can effectively capture significant patterns and relationships.

To start implementing different architectures we began with the analysis of the set of images. Since the data was coming directly from the scanner, it was decided to perform a preprocessing of them with the hope of improving the model's performance. The steps consisted of bias field correction, non-deformable registration, brain extraction, tissue and sub-cortical structures segmentation. All of the preprocessing pipeline was performed using parallel running divided into 40 constraints to achieve the best efficiency in time. The described preprocessing is depicted in Figure 3.

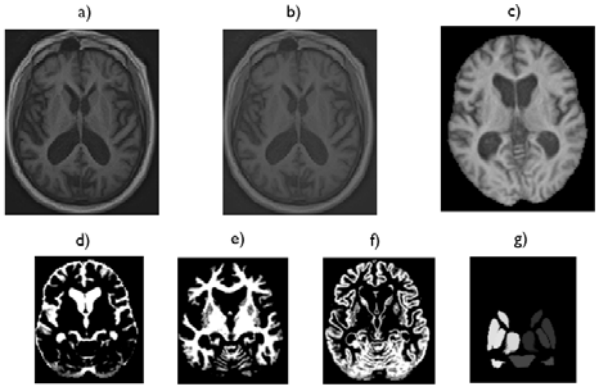


Figure 3: Visual representation of data preprocessing: a) axial slice of the original T1-weighted, b) bias field corrected, c) skull stripped registered image; d-f) CSF, WM, GM volume and g) subcortical structures

3.2.1. N4 bias field correction

The method we chose to perform bias field correction was the one presented by Tustison et al. (2010) defined in *SimpleITK* library, the current state of art algorithm.It tackles the issue of uneven intensity bias found in MRI

images. This bias can arise due to factors like irregularities in the magnetic field or artifacts during acquisition, leading to inconsistencies and distortions in the data. N4 bias field correction takes a non-parametric approach to estimate and remove this bias from the image. It achieves this by utilizing a combination of a low-frequency deformation field and a B-spline-based approximation to model the bias field. By applying this correction process, we ensure that the intensity values across different regions of the image are normalized and standardized. The output of this process is the bias field corrected image that will be used as the moving image in the registration step.

3.2.2. Non-deformable registration and skull-stripping

For the registration of the images, we used the ICBM 2009c Nonlinear Asymmetric template, (https://nist.mni.mcgill.ca/icbm-152-nonlinear-atlases-2009/), visualized in Figure 4. When comparing symmetric and asymmetric registration approaches, asymmetric atlases can capture the structural asymmetry present in the brain, such as differences in the size, shape, and location of brain regions between the left and right hemispheres. By considering these asymmetries, the atlases can provide more precise spatial alignment and mapping of brain regions. However, it is essential to ensure the correct orientation of the original images.

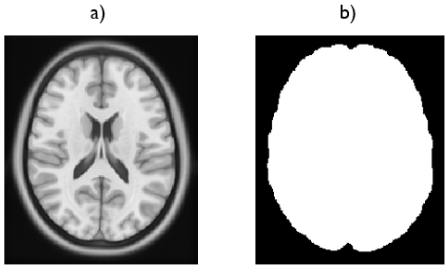


Figure 4: MNI template: a) T1 non-linear asymmetric and b) brain mask.

The preprocessing pipeline was performed using FSL library (Smith et al., 2004). This library utilizes non-deformable registration in the MNI space for skull stripping, employs the FAST algorithm for tissue segmentation, and incorporates the FIRST method for the segmentation of subcortical structures. To register the images with respect to the MNI template, we used FSL *pairreg* function, which uses the skull to maintain the global scaling of the head. This function does an affine registration but preserves the scale of the skull. In this way, we ensure that all the patients have the same intracranial volume. The non-linear registration is used for skull stripping but the image is not transformed non-linearly to the MNI space. The images are linearly transformed to the template.

3.2.3. Tissue segmentation

After having the transformed skull-stripped images into the MNI space, tissue segmentation is performed. This step is done using *fast* algorithm, Zhang et al. (2001) inside FSL library. The underlying method relies on a hidden Markov random field model and utilizes an associated Expectation-Maximization algorithm. The whole process produces 4 outputs for each patient, background, CSF, GM, and WM images.

3.2.4. Sub-cortical structures segmentation

Another feature included in fsl library is the segmentation of sub-cortical structures using *first* algorithm, (Patenaude et al., 2011). The shape and appearance model used in this approach follows multivariate Gaussian assumptions. The shape is represented by a mean shape along with modes of variation, which are essentially principal components capturing shape variations. Using these learned models, the *first* algorithm explores linear combinations of shape modes of variation to identify the most likely shape instance based on the observed intensities in a T1-weighted image.

3.3. Binary classification pipeline

We started our deep learning area using a binary classification model. In this study, the training was conducted using a small subset of cases that have extreme values for the score, resulting in the classification of the classes as “low“ and “high“ scores, assigned with 0 and 1 respectively. We adopted this classification approach as a start-up work because classifying data is generally considered easier than regression. By doing so, we aimed to demonstrate two important points:

Firstly, we wanted to establish that there are discernible patterns between brain images and the specific biological parameters we targeted. This finding would provide evidence that meaningful relationships exist between brain imaging and the variables of interest.

Secondly, we aimed to explore and test different training settings for the subsequent regression problem. By successfully achieving classification results, we could gain insights into the best approaches and configurations to be employed when training the model for regression, which involves predicting continuous numerical values rather than discrete classes.

3.3.1. 3c2d model

The model we constructed from scratch for the binary classification case is the 3D-CNN architecture presented by (Yin et al., 2023). A full overview of the training flow can be seen in Figure 5. The 3D convolutional neural network (CNN) is composed of three sequential blocks, each containing a 3D convolutional layer, a max-pooling layer, a batch normalization layer, and an optional dropout layer.

Following the convolutional blocks, the fourth block comprises a global average pooling layer specifically

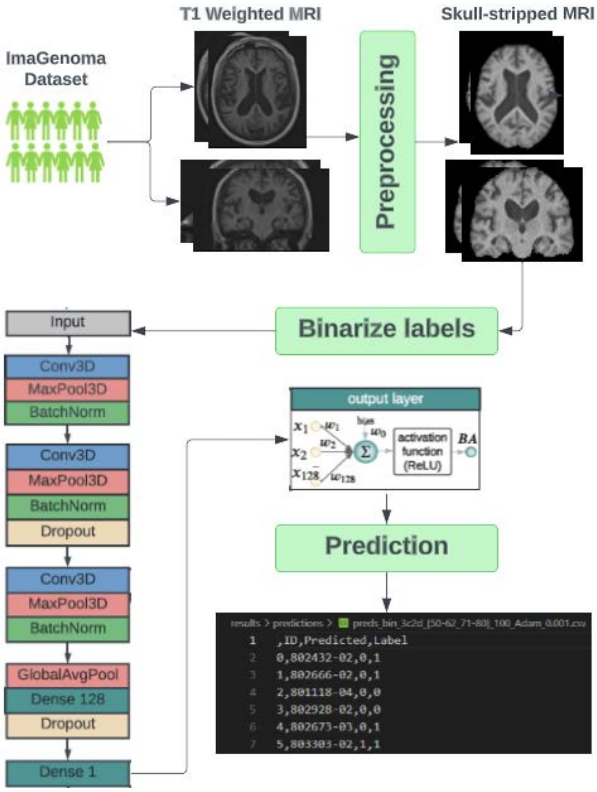


Figure 5: Working flow of the binary classification pipeline.

designed for 3D data. This pooling layer reduces the feature map of size 18 x 18 x 18 x 128 to a pooled representation of 128 x 1. Subsequently, the pooled features are passed through a dense layer and a dropout layer with a dropout rate of 0.3. In the original design, the output dense layer had one output neuron to estimate BA using regression. However, for our adaptation, we modified the number of output neurons to 2 since we are now dealing with binary classification. We named this model '3c2d' since it is constructed by 3 convolutional layers and 2 dense layers.

3.3.2. Loss functions

Mean square error (MSE) loss

It is defined as:

$$MSE = \frac{1}{N} * \sum_{i}^{N} (Y_{\mathrm{i}} - F_{\mathrm{Xi}}) \tag{1}$$

where:

- N is the total number of subjects
- Y_i are the true ages.
- F_{Xi} represents the neural network that outputs the predicted age directly.

Cross-entropy (CE) loss

The cross-entropy loss function calculates the average negative log-likelihood of the predicted age class probabilities. It penalizes the model for deviating from the true age labels, encouraging it to learn accurate age predictions. The equation for cross-entropy loss is as follows:

$$\mathrm{CE} = -\sum_{i}^{N} t_i \cdot \log Ft_i \tag{2}$$

where:

- t_i is the target value for i_{th} index.
- Ft_i is the i_{th} scalar value of the model output.

3.3.3. Training and testing

ImaGenoma dataset was divided into 60/20/20 for training, validation, and test set respectively ensuring a balanced splitting. In the training phase, we employed a Stochastic Gradient Descent (SGD) and an Adam optimizer. The interval of the age we chose to work on was [50, 62] and [71, 80]. This was due to the fact that these intervals present the edges of our distribution of ages. Also by choosing these extremes, we avoided the unbalancing problem.

In order to be compatible with the classification architecture, we "binarized" the true ages. This step is executed if the length of the interval is not 0 and the binarize option is set to 'true'. The age values corresponding to the lower bound of the interval are set to 0 and the other ones are set to 1. In this way we end up working with binarized age labels as shown in Figure 6.

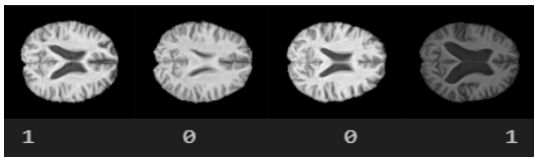


Figure 6: Examples of brain MRI with binarized age scores.

Different combinations of the parameters were conducted whereas the best resulting one was using MSE loss and SGD optimizer. The learning rate was set to 0.001. A detailed explanation of the results can be found in Section 4.1.

3.4. Regression deep learning model

In contrast to the binary classification pipeline, in this section, a deep learning pipeline for regression is exploited. For this approach, we followed the best-performing architecture published by Peng et al. (2021b), SFCN. For the authors, SFCN provided the best result (MAE 2.14) in predicting age from MRI images, trained with Uk Biobank data. We tested this model by adapting their architecture using our dataset.

3.4.1. SFCN model

The SFCN model is structured with seven convolution blocks. The first five blocks follow a consistent pattern: a 3D convolution layer, a batch normalization layer, a max pooling layer, and a ReLU activation layer. What sets this architecture apart is that it downsamples the input after each convolution layer. This means that as the model progresses through the layers, the spatial dimension of the data reduces rapidly. The input size is reduced from 160x192x160 to 5x6x5 (voxels) in the first 5 blocks.

The sixth block of the architecture follows a similar structure but with some variations. Unlike the previous blocks, it does not include a max pooling layer. Instead, it incorporates a 1x1x1 3D convolution layer, which introduces non-linearity while preserving the spatial dimensions of the feature map. Following the convolutional layer, a resulting feature map of size 5x6x5 is obtained. This feature map is then processed by an average pooling layer, which reduces its spatial dimensions while retaining the important features. Subsequently, the pooled feature map is fed into an output layer through a linear transformation, achieved by a fully connected layer. The input size remains consistent for both T1 non-linearly registered brains and linearly registered brains. The dimensions are 160x192x160 voxels. The input data for the deep neural network model was brain extracted, bias-corrected, and linearly registered to MNI152 standard space. The head size of subjects is normalized as a result of the linear registration.

3.4.2. Training and testing

In the training phase, the authors employed a Stochastic Gradient Descent (SGD) optimizer, following the work by Sutskever et al. (2013), to train their model on the UK Biobank (UKB) dataset. The objective was to minimize the Kullback-Leibler divergence loss function between the predicted probability and a Gaussian distribution. In this distribution, the mean was set as the true age of each training subject, while the standard deviation (sigma) was fixed at 1 year specifically for the UKB dataset. When using PAC 2019 dataset for testing their approach, sigma was set to 2 denoting a 2-year age interval. The L2 weight decay coefficient was 0.001. The batch size was 8. The learning rate for the SGD optimizer was initialized as 0.01, then multiplied by 0.3 every 30 epochs unless otherwise specified.

3.4.3. Model output and loss function

The predicted probability that the subject's age falls into a one-year age interval is represented by 40 digits in the output layer. To calculate the final prediction, each age bin's weighted average is calculated:

$$\text{pred} = \sum_{i}^{40} x_i \cdot \text{ age }_i \tag{3}$$

where x_i stands for the probability predicted for the i^{th} age bin and age_i stands for the bin centre in the age interval.

In this particular approach, the age label is not considered as a single precise value, but rather as a range represented by a discretized Gaussian probability distribution centered around the true age. Similarly, the model's output is also in the form of a probability distribution (Figure 7).

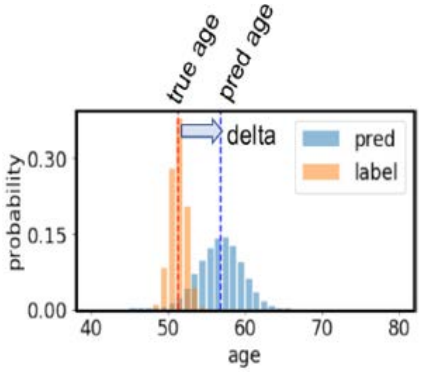


Figure 7: An example of soft labels and output propabilities (Peng et al., 2021b).

For the loss function, the Kullback-Leibler divergence, also known as relative entropy, is used. It represents a measure of the difference between two probability distributions. In the context of age prediction, the KL divergence can be used to quantify the dissimilarity

between the predicted age distribution and the true age distribution. The equation for Kullback-Leibler divergence between two probability distributions P and Q is given by:

$$KL(P||Q) = \sum \frac{P(x) * \log(P(x))}{Q(x)} \quad (4)$$

where:

- P(x) represents the probability of age x in the true age distribution.

- Q(x) represents the probability of age x in the predicted age distribution.

The KL divergence measures the average number of additional bits needed to represent data from the true distribution P when using a code based on the predicted distribution Q. It is a non-negative value, where a lower KL divergence indicates a closer match between the two distributions.

In the context of age prediction, minimizing the KL divergence can be a useful objective to optimize the prediction model and improve the alignment between the predicted age distribution and the true age distribution. An example of how the prediction and age labels are distributed is shown in Figure 8.

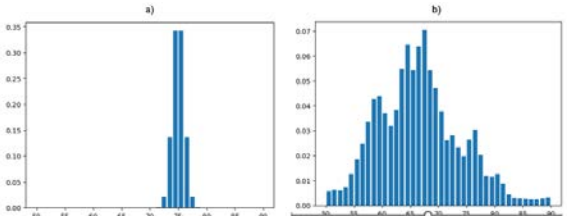


Figure 8: Example of a) soft label and b) prediction distribution for one of the subjects.

3.4.4. Architecture Adaption, Optimiser Choice and Training

The working pipeline of the model can be seen in Figure 9. To be compatible with the model requirements we adapted some changes to our dataset. Our input data was originally 193x229x193 voxel size. In order to match the input size the model is expecting, 2 different approaches were followed:

1. Change the spatial size to 256x256x256 padding with 0s.
2. Cropping the field of view (FOV) to 160x192x160.

For the first approach, we used *SpatialPad* function defined in MONAI library (Cardoso et al., 2022). This transformation performs padding to the data, symmetric for all sides or all on one side for each dimension. In our case, we padded the images to 256x256x256 with 0s in all sides. We chose this size because of how the convolution layers work. In order to make the convolutions work, the input shape should be divisible by powers of

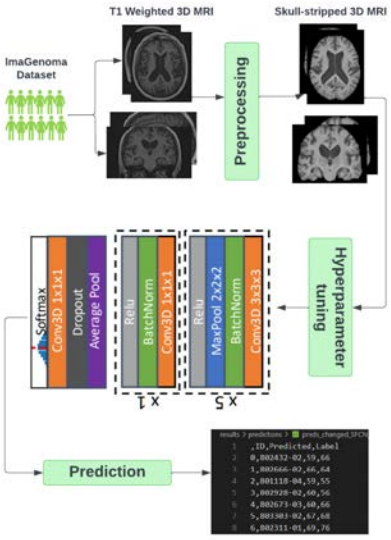


Figure 9: Overview of the deep learning regression pipeline.

2, which means, $2^8 = 256$. Figure 10 shows an example image after applying *SpatialPad*. We can see that the only change is in the background which now is bigger in size.

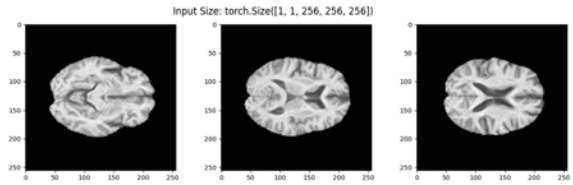


Figure 10: Visualization of one image at different slices after SpatialPad().

The first step applied for the second approach was to crop the MNI9c registered data to match the dimensions of the MNI6th template, (http://bic.mni.mcgill.ca/ServicesAtlases/ICBM152NLin6) which was used for the registration part of Uk Biobank data and removed offset due to odd mni9c dimensions by subtracting each dimension by 1. After that, the data is normalized by diving it with the mean and as a last step, cropped the FOV to match the expected input of the CNN to 160x192x160. This approach achieved better results. This can be due to the normalization part, where dividing an image that has more 0 values will give us a smaller intensity value. Figure 11 refers to an example image after cropping it to 160x192x160.

The first trials made were using SGD optimizer and the same parameters the authors used in their architectures. After trying Adam optimizer, we observed that for our dataset Adam worked better (a detailed explanation can be found in section 4.2). For this reason, we continued all of our experiments using Adam instead of SGD optimizer.

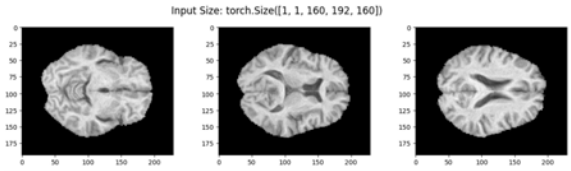


Figure 11: Visualization of one image at different slices after cropping to 160x192x160.

Several experiments were employed for hyper-parameter tuning during the experimentation process. These strategies included fine-tuning all layers using pre-saved weights, inference on the test set, fine-tuning only the classifier block of the model, initializing weights using the Xavier initialization technique, (Kumar, 2017), and training from scratch. Another interesting experiment was to fine-tune only the batch normalization layers as suggested by (Kanavati and Tsuneki, 2021). As a last trial, we did a set of freezing and unfreezing of the blocks. To explore the impact of various hyper-parameters, all the experiments were conducted, each involving different values of the learning rate, batch size, and optimizer.

3.5. Metrics analysis

3.5.1. Binary classification problem

Accuracy

In the context of the binary classification pipeline used in age prediction, accuracy is a metric that measures the overall correctness of the model's predictions. It quantifies the proportion of correctly classified instances out of the total number of instances in the dataset. The equation for accuracy is given by:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

where:

-TP (True Positive) is the number of correctly predicted positive instances.

-TN (True Negative) is the number of correctly predicted negative instances.

-FP (False Positive) is the number of instances falsely predicted as positive (Type I error).

-FN (False Negative) is the number of instances falsely predicted as negative (Type II error).

For our model, we did an derived version of the accuracy. The accuracy is calculated as:

$$ACC = \frac{\text{correct}}{\text{number of patients}} \tag{6}$$

In our case if $|P - T| < 1$, correct variable is increased by 1. P denotes predicted and T stands for true.

Sensitivity

Sensitivity (also known as recall) is a metric that evaluates the model's ability to correctly identify positive instances from the actual positive instances in the dataset. It quantifies the proportion of true positive predictions out of the total number of actual positive instances. The equation for sensitivity is given by:

$$SENS = \frac{TP}{TP + FN} \tag{7}$$

where:

-TP: P=1 & T=1

-FN: P=0 & T=1

-TN: P=0 & T=0

-FP: P=1 & T=0

Specificity

On the other hand, specificity identifies negative instances from the actual negative instances in the dataset. It quantifies the proportion of true negative predictions out of the total number of actual negative instances. Specificity is calculated by:

$$SPEC = \frac{TN}{TN + FP} \tag{8}$$

3.5.2. Regression problem

Mean absolute error

In the evaluation process of the deep learning model, we used the mean absolute error (MAE) metric and r-squared coefficient. MAE provides a measure of the average absolute difference between the predicted age and the ground truth age labels, typically expressed in years. The equation for mean absolute error (MAE) is given by:

$$MAE = \frac{1}{N} * \sum \left| x_{\text{true}} - x_{\text{pred}} \right| \tag{9}$$

where:

-N is the total number of samples.

-x_{true} represents the true age labels.

-x_{pred} represents the predicted age values.

In our implementation, we calculated MAE as:

$$MAE = \frac{\text{absolute\_errors}}{\text{number of patients}} \tag{10}$$

where absolute_errors += $|P - T|$.

r2 coefficient

The deep learning model was also evaluated using r2 coefficient. Also known as the coefficient of determination, it measures the proportion of the variance in the dependent variable (age) that can be explained by the independent variables (features) in the model. A higher r2 score suggests that the model successfully explains a larger proportion of the variance in age, indicating better predictive performance. r2 coefficient is calculated as:

$$r2 = 1 - \frac{\text{SSR}}{\text{SST}} \tag{11}$$

where:

-SSR (Sum of Squared Residuals) is the sum of the squared differences between the true age values and the predicted ones.

-SST (Total Sum of Squares) is the sum of the squared differences between the true age values and their mean.

Figure 12a) refers to a perfect prediction matching the diagonal. Meanwhile in 12b) we can see one of the cases out of our trials.

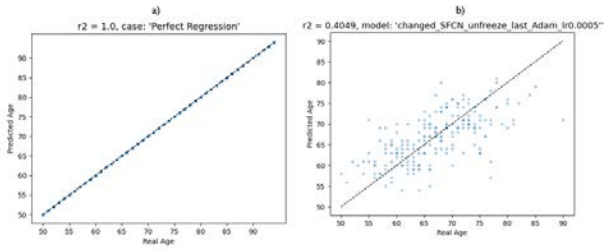


Figure 12: Different cases of r2 coefficient.

3.6. Implementation Details

This thesis was implemented using Jupyter Notebook and Python programming language. Additional libraries that were employed include torch, os, pandas, nibabel and matplotlib. The visualization of the 3D volumes was done in ITK Snap (Yushkevich et al., 2016). The bias field correction was done using simpleitk's implementation of N4 algorithm (Tustison et al., 2010). Image registration and skull stripping, segmentation of tissues, and subcortical structures were applied using fsl_anat package introduced by (Smith et al., 2004). The deep learning models were implemented using PyTorch (Paszke et al., 2019). 2 NVIDIA-SMI GPUs were used, 12G and 24G respectively whereas CUDA Version was 12.1.

4. Results

In accordance with the aforementioned architectures, the results section will be organized to include the indicated experiments.

4.1. The performance of 3c2d in ImaGenoma dataset

The first experiment using 3c2d model was choosing MSE as the loss function and learning rate = 0.01, as the authors suggested in their publication. To check the impact of the optimizer the same experiment but with a different optimizer was repeated.

Taking into consideration that SGD gave better results, another experiment with a different learning rate value was conducted. For comparison purposes, CE loss was tried together with Adam optimizer instead of SGD, keeping the same value of the learning rate. A full description of the results and the corresponding parameters can be found in Table 3.

To conclude we can say that the combination of the 3c2d model architecture, SGD optimizer with a learning rate of 0.001, and MSE loss function demonstrated superior performance compared to the other results.

4.2. The performance of SFCN on ImaGenoma dataset

4.2.1. Initial experiments

The SFCN (Peng et al. (2021b)) architecture was run using a series of experiments, exploring different parameters and methods. A diverse range of methods were utilized, focusing on the input images obtained through padding to size 256x256x256 using *SpatialPad()* and cropping to 160x192x169 dimensions. A series of initial experiments were run using both input data, like finetuning all layer using presaved weights, inferencing directly on test set, finetuning only the last block of the architecture, and training from scratch using xavier initialization method. The impact of the input data as a result of padding and cropping (detailed explanation in Section 3.4.4, in these experiments, can be observed in Table 4. We can see from the table that the model gave better results (MAE = 5.24 years) in the case when only the classifier block of the architecture was finetuned. This emphasizes the fact that by fine-tuning only the last block of the model, it concentrates on learning task-specific features without modifying the earlier representations that have already been learned. This way it prevents overfitting. All the above experiments were conducted using SGD optimizer since it gave the best results for the authors.

The span of the predicted values was improved and larger when using cropped data. The number of points passing through the regression line was increased, i.e. the number of correctly predicted values was higher.

Based on Table 4, different values of parameters were tried using both types of input data for both of the methods. This includes different values of learning rate and optimizers. These results can be found in Table 5.

Deriving from the results of the above table, we observed an improvement in terms of MAE (MAE = 4.52 years) and r2 (r2 = 0.2924) when using Adam as an optimizer and working with the input data cropped to 160x190x160. This can be due to the fact that the Adam optimizer is known for its adaptive learning rate mechanism, leading to faster convergence and improved performance. Furthermore, cropping the input data to a smaller size eliminates unnecessary background or empty space that may not contribute significantly to the model. Consequently, the model becomes more effective in capturing the essential information necessary for accurate age prediction. Based on this, all the next experiments will be conducted keeping unchanged these parameters.

4.2.2. Finetuning and choosing the best model

Taking inspiration from Table 5, the best-performing model was found to be finetuning only the classification

Table 3: The performance of 3c2d model on ImaGenoma Dataset.

Model	Epochs	Learning rate	Loss function	Batch size	Patience	Accuracy	Sensitivity	Specificity
3c2d(Adam)	300	0.01	MSE	8	30	0.48	0.55	0.34
3c2d(Adam)	100	0.001	CE	4	10	0.57	0.62	0.29
3c2d(SGD)	300	0.01	MSE	8	30	0.59	0.63	0.31
3c2d(SGD)	100	0.001	MSE	4	10	**0.76**	**0.79**	**0.16**

Table 4: Comparison of the results between different input data.

Method	SpatialPad()		Cropping	
	MAE(years)	r2	MAE (years)	r2
Finetune all layers using pre-saved weights	5.78	0.0095	5.88	0.0167
Inference on test set	6.19	0.0993	6.57	0.1800
Finetune only the classification block	7.73	0.8619	**5.24**	**0.0770**
Initialize weights using xavier and train from scratch	6.10	0.8207	5.86	0.0168

Table 5: Hyperparameters tuning for finetune all layers model.

	lr	MAE(years)	r2	Optimizer
SpatialPad()				
	0.001	5.78	0.0095	SGD
	0.01	8.11	0.9271	SGD
	0.001	5.88	0.0173	Adam
	0.0001	7.73	0.8619	SGD
	0.0001	7.14	0.6011	Adam
Cropping				
	0.001	5.88	0.0167	SGD
	0.01	5.93	0.0218	SGD
	0.001	5.85	0.8255	Adam
	0.0001	5.24	0.0770	SGD
	0.0001	**4.52**	**0.2924**	Adam

(last) block. Building on this result, we further investigated the influence of different learning rates using the Adam optimizer. Our experiments, presented in Figure 13, provide insights into the impact of various learning rates on the overall performance of the model. The best results were met when using a learning rate of 0.00001, MAE = 3.33 years and r2 = 0.6713.

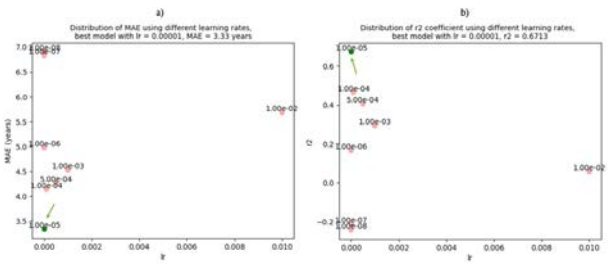


Figure 13: Distribution of a) MAE and b) r2 using different learning rates, best model with lr = 1e-05, MAE = 3.33 years and r2 = 0.6713.

All the experiments were conducted with a batch size of 1 to accommodate the size of the MRI volume and GPU limitations. However, to get the best out of the model, a batch size of 3 was utilized. Changing the batch size did not lead to an improvement in the model's performance; in fact, it resulted in a decay of the model's performance. The experiment with a larger batch size showed worse results compared to the experiments conducted with a batch size of 1. This can be due to the fact that when employing a larger batch size, the model learns from a combination of samples that may possess diverse features and characteristics. The increased variability within the batch can pose challenges for the model in extracting meaningful patterns and achieving effective generalization. On the other hand, a smaller batch size of 1 enables the model to concentrate on individual samples and their unique characteristics. Furthermore, using a larger batch size can introduce noise in the gradient estimation process due to the combination of multiple samples. This noise can have a negative impact on the accuracy of parameter updates during training, potentially impeding convergence or resulting in sub-optimal solutions.

4.2.3. Testing other methods

We explored a recent approach inspired by the work of Kanavati and Tsuneki (2021), which involved fine-tuning only the batch normalization layers. The researchers discovered that selectively fine-tuning the trainable weights of the batch normalization layers yielded comparable performance to fine-tuning all weights while achieving faster convergence. In another trial, we employed a strategy of freezing and unfreezing steps. Initially, we fine-tuned only the classification block of the SFCN model for 10 epochs. Subsequently, the feature extraction part block was unfrozen for 3 epochs, and as a last step, we re-fine-tuned the classification block for an additional 10 epochs. The objective behind this approach was to train the feature extraction part for a few epochs, enabling the model to acquire specific data-related features and prevent strong overfitting.

The results obtained were better with a value of 0.17 and worse with a value of 0.11 for finetuning batch normalization layers and set of freezing and unfreezing approaches respectively. Their corresponding Pearson correlation coefficient is visualized in Figure 14 and shown in Table 6.

In summary, among the various trials conducted, the

Table 6: MAE and r2 for training only the batch normalization layers and freezing/unfreezing the blocks of SFCN architecture.

Method	MAE (years)	r2
Unfreeze batch normalization layers	**4.35**	**0.4128**
Combination of freeze and unfreeze of layers	4.63	0.3281

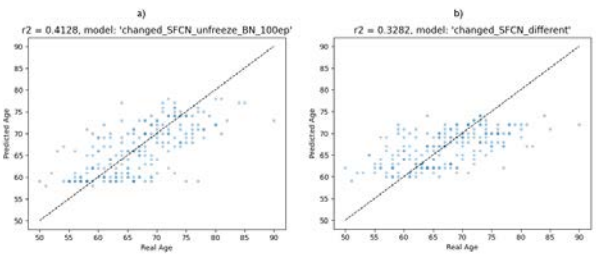


Figure 14: r2 plot a)for training only the batch normalization layers and b) freezing/unfreezing the blocks of SFCN architecture.

best-performing model was achieved by fine-tuning the last block of the architecture using a learning rate of 0.00001 and Adam as the optimizer. This model resulted in an MAE = 3.33 years and r2 = 0.6713. This particular configuration yielded superior results compared to other experiments. A detailed overview of the parameter tuning process and the methods employed can be found in Table 7, providing detailed insights into the performance and effectiveness of each approach.

4.2.4. Comparison with the state of the art

Deriving from section 2 this part will compare our results with the ones achieved by previous authors holding the state of the art. As it can be seen from Table 1, our approach lies in the deep learning area with an MAE equal to the one achieved by Inception V1. This approach utilized a priori information taking GM and in contrast, we used the whole volume to train the models. Our approach did not outperform the original paper that used SFCN model. This may be due to the dataset size, hyperparameter tuning or presaved weights. An full comparison of the results and the place where our approach takes place only in deep learning area can be seen in Table 8.

5. Discussion

The models presented in this thesis focus on predicting age based on brain MRI scans, particularly emphasizing the change of brain tissues observed in healthy individuals. The modality employed for image acquisition is T1-weighted imaging. By comparing an individual's predicted brain age to their chronological age, it becomes possible to detect early signs of accelerated or delayed brain aging, indicating potential neurodegenerative conditions or cognitive impairments.

Primarily, a binary classification architecture was constructed from scratch with the goal of differencing "high" and "low" values. By constructing the binary classification architecture from scratch, it was possible to have full control over the model's architecture, including the choice of layers, activation functions, and connectivity patterns. The age labels were binarized before feeding to the network and the output layers was changed to 2 neurons indicating the binary classification task. Two different loss functions were used, MSE and CE. We also tested the impact of the optimizer on the model by hyperparameter tuning between Adam and SGD optimizer.

The best model resulted to be when using SGD as an optimizer, a learning rate of 0.001, and MSE as a loss function. Having a learning rate of 0.001 indicates relatively small steps, which can help the model converge gradually and avoid overshooting the optimal parameter values. The loss function and the optimizer also contributed effectively to capturing the discrepancies between the predicted probabilities and the true binary labels.

Secondly, another deep learning model named SFCN was implemented. Fine-tuning all layers, inferencing the test set directly, training from scratch, and finetuning only the classification block were the first experiments. For this models, we used the parameters suggested by the authors in their paper and used 2 different input data: padding with 0 the input to 256x256x256 and cropping to 160x190x160. The best model resulted to be the one where you fine-tune the last block of the architecture. After hyperparameter tuning, the best parameters resulted to be Adam optimizer with a learning rate of 0.00001 and input data cropped to the model requirements. This resulted also in the best model of our experiments. To furthermore investigate and improve the results we did two other trials based on these parameters which were unfreezing and updating the weights of the batch normalization layers only and doing a set of freezing and unfreezing of the layers of the architecture. However, none of them did not surpass the performance of the best model.

6. Limitations

While the model showcased promising and consistent outcomes, there are still several limitations that could be addressed through future research. It's worth noting that the model encounters challenges and tends to overfit in the age range of 60-70 years due to the imbalanced distribution of the dataset, with a majority of cases falling

Table 7: The performance of different variations of SFCN model on ImaGenoma Dataset. *epochs = 100, loss = KL-Div, patience = 10*

Method	Learning rate	Optimizer	Batch size	MAE(years)	r2
Finetune all layers using pre-saved weights	0.001	SGD	1	5.88	0.0167
Inference on test set	0.001	SGD	1	6.57	0.1800
Finetune only the classification block	0.001	SGD	1	5.24	0.0770
Initialize weights using xavier and train from scratch	0.001	SGD	1	5.86	0.0168
Finetune only the classification block	0.00001	Adam	1	**3.33**	**0.6713**
Unfreeze batch normalization layers	0.00001	Adam	1	4.35	0.4128
Combination of freeze and unfreeze of the layers	0.00001	Adam	1	6.63	0.3821
Finetune only the classification block	0.00001	Adam	3	5.05	0.0566

Table 8: Overview of the state-of-the-art and our method in deep learning area.

Methodology	Input	Author	Dataset	MAE (years)
SFCN	3D Volume	**Deep Learning** Peng et al. (2021b)	UK Biobank PAC 2019	2.14 2.9
3D Resnet-34	3D Volume	**Deep Learning** Zhang et al. (2023)	UK Biobank OASIS ABIDE	2.55
SFCN	3D Volume	**Deep Learning** Our Method	ImaGenoma	**3.33**
Inception V1	Priori info: GM	**Deep Learning** Couvy-Duchesne et al. (2020)	PAC 2019	3.33
3D Resnet-18	3D Volume	**Deep Learning** Barbano et al. (2022)	OpenBHB	3.76

within that interval. Furthermore, the dataset used for training the model is relatively small in size.

When employing complementary task learning, particularly in the regression aspect, it was observed fluctuations in the KL loss during the training process. Unfortunately, in many instances, the model tends to predict values within a narrow range, resulting in significant overfitting to the training data. As a consequence, the mean absolute error (MAE) was high, as neither the selection of an appropriate loss function nor the optimization method was properly optimized.

7. Future work

In order to improve the prediction results, enhancing the registration process and improving skull stripping techniques could potentially enhance the model's performance. In addition, including prior information related to WM and/or GM can improve the pipeline. Introducing data augmentation with appropriate transformations could lead to a more robust method.

Further optimization, includes fine-tuning the number of encoding/decoding blocks and convolutional layers, as well as incorporating batch normalization techniques. Exploring different architectures can also contribute to achieving improved results.

Finally, incorporating multi-modal data fusion, and developing interpretable models would have a great significance for better understanding and explainability.

8. Conclusions

In conclusion, the utilization of MRI-based brain age prediction has yielded promising outcomes in estimating an individual's brain age. This study incorporated two pipelines, specifically a binary classification model and a SFCN (Simple Fully Convolutional Network) model, to forecast brain age based on MRI data.

The binary classification model effectively categorized individuals into two groups: younger or older, based on their brain age. Although this approach provided valuable insights regarding relative age differences, it lacked the capability to offer a continuous and more precise estimation of an individual's brain age.

In contrast, the SFCN model showcased better performance in accurately predicting brain age. Notably, the model achieved optimal results by fine-tuning only the classifier block, suggesting that the lower-level feature extraction layers of the SFCN model already captured relevant patterns and representations from the MRI data effectively. By focusing the finetuning process on the classifier block, the model can effectively adapt to specific brain age prediction tasks without sacrificing the learned representations from the lower-level layers.

In summary, the integration of MRI-based brain age prediction models, especially the successful implementation of the SFCN model, not only advances our understanding of brain aging but also offers distinct clinical benefits. This technology has the potential to revolutionize clinical practice by providing valuable insights into brain health, aiding in early detection, and facilitating targeted interventions to optimize brain function and promote healthy aging.

Acknowledgments

I would like to thank my supervisors, Xavier Llado and Arnau Oliver for their invaluable support and continuous guidance throughout the timeline of this thesis. Their assistance and support have been essential in the successful completion of this research project. To Dr. Oren Contreras of the Hospital and Dr. Josep Trueta for sharing the dataset. Last but not least, I would like to thank all the people that became part of my MAIA experience.

References

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N.K., Andersson, J.L., Griffanti, L., Douaud, G., Sotiropoulos, S.N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., et al., 2018. Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. Neuroimage 166, 400–424.

Baecker, L., Dafflon, J., Da Costa, P.F., Garcia-Dias, R., Vieira, S., Scarpazza, C., Calhoun, V.D., Sato, J.R., Mechelli, A., Pinaya, W.H., 2021. Brain age prediction: A comparison between machine learning models using region-and voxel-based morphometric data. Human brain mapping 42, 2332–2346.

Barbano, C.A., Dufumier, B., Duchesnay, E., Grangetto, M., Gori, P., 2022. Contrastive learning for regression in multi-site brain age prediction. arXiv preprint arXiv:2211.08326 .

Baxi, M., Di Biase, M.A., Lyall, A.E., Cetin-Karayumak, S., Seitz, J., Ning, L., Makris, N., Rosene, D., Kubicki, M., Rathi, Y., 2020. Quantifying genetic and environmental influence on gray matter microstructure using diffusion mri. Cerebral Cortex 30, 6191–6205.

Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al., 2022. Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 .

Cole, Ritchie, Bastin, S.M, 2017. Brain age predicts mortality. URL: `https://www.nature.com/articles/mp201762`.

Cole, J.H., Franke, K., 2017. Predicting age using neuroimaging: innovative brain ageing biomarkers. Trends in neurosciences 40, 681–690.

Cole, J.H., Franke, K., the Alzheimer's Disease Neuroimaging Initiative, 2010. Predicting age using neuroimaging: Innovative brain aging biomarkers. Trends in Neurosciences 33, 628–637.

Cole, J.H., Poudel, R.P.K., Tsagkrasoulis, D., Caan, M.W.A., Steves, C., Spector, T.D., Montana, G., 2019. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. NeuroImage 163, 115–124.

Couvy-Duchesne, B., Faouzi, J., Martin, B., Thibeau-Sutre, E., Wild, A., Ansart, M., Durrleman, S., Dormont, D., Burgos, N., Colliot, O., 2020. Ensemble learning of convolutional neural network, support vector machine, and best linear unbiased predictor for brain age prediction: Aramis contribution to the predictive analytics competition 2019 challenge. Frontiers in Psychiatry 11, 593336.

Da Costa, P.F., Dafflon, J., Pinaya, W.H., 2020. Brain-age prediction using shallow machine learning: predictive analytics competition 2019. Frontiers in psychiatry 11, 604478.

De Fauw, J., Ledsam, J.R., Romera-Paredes, B., et al., 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Medicine 24, 1342–1350. URL: `https://doi.org/10.1038/s41591-018-0107-6`, doi:10.1038/s41591-018-0107-6.

Fisher, E., Lee, J.C., Nakamura, K., Rudick, R.A., 2008. Gray matter atrophy in multiple sclerosis: a longitudinal study. Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society 64, 255–265.

Franke, K., Ziegler, G., Klöppel, S., Gaser, C., the Alzheimer's Disease Neuroimaging Initiative, 2009. Estimating the age of healthy subjects from t1-weighted mri scans using kernel methods: Exploring the influence of various parameters. NeuroImage 50, 883–892.

Good, C.D., Johnsrude, I.S., Ashburner, J., Henson, R.N., Friston, K.J., Frackowiak, R.S., 2001. A voxel-based morphometric study of ageing in 465 normal adult human brains. Neuroimage 14, 21–36.

Jönemo, J., Akbar, M.U., Kämpe, R., Hamilton, J.P., Eklund, A., 2022. Efficient brain age prediction from 3d mri volumes using 2d projections. arXiv preprint arXiv:2211.05762 .

Kanavati, F., Tsuneki, M., 2021. Partial transfusion: on the expressive influence of trainable batch norm parameters for transfer learning, in: Medical Imaging with Deep Learning, PMLR. pp. 338–353.

Kaufmann, T., Alnæs, D., Brandt, C.L., Doan, N.T., Kauppi, K., Bettella, F., the Alzheimer's Disease Neuroimaging Initiative, the NeuroCHARGE Working Group, the Pediatric Imaging, Neurocognition, and Genetics Study, the IMAGEN Consortium, the Alzheimer's Disease Metabolomics Consortium, the Alzheimer's Disease Neuroimaging Initiative, Westlye, L.T., 2017. Task modulations and clinical manifestations in the brain functional connectome in 1615 fmri datasets. NeuroImage 147, 243–252.

Koutsouleris, N., Davatzikos, C., Bottlender, R., Patschurek-Kliche, K., Scheuerecker, J., Decker, P., Gaser, C., the German Research Network on Schizophrenia, the Alzheimer's Disease Neuroimaging Initiative, Meisenzahl, E.M., 2012. Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. Archives of General Psychiatry 69, 220–229.

Kumar, S.K., 2017. On weight initialization in deep neural networks. arXiv preprint arXiv:1704.08863 .

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. nature 521, 436–444.

Liem, F., Varoquaux, G., Kynast, J., Beyer, F., Kharabian, S., Huntenburg, J.M., 2015. Predicting brain-age from multimodal imaging data captures cognitive impairment. NeuroImage 148, 179–188.

Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L., et al., 2016. Multimodal population brain imaging in the uk biobank prospective epidemiological study. Nature neuroscience 19, 1523–1536.

Mitchell, T.M., et al., 2007. Machine learning. volume 1. McGraw-hill New York.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems 32.

Patenaude, B., Smith, S.M., Kennedy, D.N., Jenkinson, M., 2011. A bayesian model of shape and appearance for subcortical brain segmentation. Neuroimage 56, 907–922.

Peng, H., Gong, W., Beckmann, C.F., Vedaldi, A., Smith, S.M., 2021a. Accurate brain age prediction with lightweight deep neural networks. Medical Image Analysis 68, 101871. doi:https://doi.org/10.1016/j.media.2020.101871.

Peng, H., Gong, W., Beckmann, C.F., Vedaldi, A., Smith, S.M., 2021b. Accurate brain age prediction with lightweight deep neural networks. Medical image analysis 68, 101871.

Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., et al., 2004. Advances in functional and structural mr image analysis and implementation as fsl. Neuroimage 23, S208–S219.

Sutskever, I., Martens, J., Dahl, G., Hinton, G., 2013. On the importance of initialization and momentum in deep learning, in: International conference on machine learning, PMLR. pp. 1139–1147.

Taki, Y., Kinomura, S., Sato, K., Goto, R., Kawashima, R., Fukuda, H., 2011. A longitudinal study of gray matter volume decline with age and modifying factors. Neurobiology of Aging 32, 907–915.

Thompson, P.M., Hayashi, K.M., De Zubicaray, G., Janke, A.L., Rose, S.E., Semple, J., Herman, D., Hong, M.S., Dittmer, S.S., Doddrell, D.M., et al., 2003. Dynamics of gray matter loss in

alzheimer's disease. Journal of neuroscience 23, 994–1005.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4itk: improved n3 bias correction. IEEE transactions on medical imaging 29, 1310–1320.

Yin, C., Imms, P., Cheng, M., Amgalan, A., Chowdhury, N.F., Massett, R.J., Chaudhari, N.N., Chen, X., Thompson, P.M., Bogdan, P., et al., 2023. Anatomically interpretable deep learning of brain age captures domain-specific cognitive impairment. Proceedings of the National Academy of Sciences 120, e2214634120.

Yushkevich, P.A., Gao, Y., Gerig, G., 2016. Itk-snap: An interactive tool for semi-automatic segmentation of multi-modality biomedical images, in: 2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC), IEEE. pp. 3342–3345.

Zhang, B., Zhang, S., Feng, J., Zhang, S., 2023. Age-level bias correction in brain age prediction. NeuroImage: Clinical , 103319.

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. IEEE transactions on medical imaging 20, 45–57.

Medical Imaging and Applications

Master Thesis, June 2023

Binary classification and detection of large-vessel occlusions in acute ischemic stroke

Paola Martínez Arias[a], Uma Maria Lal-Trehan Estrada[a], Mikel Terceño[b], Luca Giancardo[c], Arnau Oliver[a], Xavier Lladó[a]

[a] *VICOROB Research Group, Girona, Spain*
[b] *Hospital Dr. Josep Trueta, Girona, Spain*
[c] *UTHealth, Houston, TX, USA*

Abstract

A Large Vessel Occlusions (LVO) is a specific type of acute ischemic stroke. It refers to the complete or partial blockage of one of the brain's major arteries. The established treatment for it is an endovasculas thrombectomy (EVT) which is more effective within 6 hours after the symptoms. Consequently, this time sensitive LVO identification is crucial to improve the patient outcomes after the episode. This work is focused on mainly two tasks: binary classification of CTA images as LVO present or absent, and detection of the exact 3D localization of the occlusion through a bounding box. We work with two datasets: data from the IACTA-EST-2023 challenge that will be used for the classification problem, and data from the Dr. Josep Trueta Hospital, in Girona, Spain. We will use deep learning techniques (DL) for both tasks. For the classification part, we propose a comparative analysis between the use of symmetry of the brain with different strategies, against not using symmetry information when training the DL model. Results show that the use of symmetry information improves the results. The top performing experiments achieved 84% in accuracy, 88% in specificity, 81% in sensitivity and an AUC of 0.895 on the test set, and 77% accuracy of inference on the hospital dataset, that was not used for training. For the detection part, we propose the use of nnDetection framework, which has not been used for this purpose before. The model is trained with data from the hospital, and obtains promising results in both anterior and posterior circulation occlusions. Results of the detector show a 97% of sensitivity in the test cases, with approximately 0.15 FPpI. To the best of our knowledge, this is the first detection proposal for automatic occlusion detection using only CTA images that is not part of commercial software with undisclosed algorithms.

Keywords: Ischemic Stroke, LVO, Deep Learning, Classification, Detection

1. Introduction

Acute Ischemic Stroke (AIS) denotes a medical condition characterized by the blocking of a cerebral artery, resulting in the sudden interruption of the blood supply and subsequent damage to distinct cerebral regions. It can occur because of a blood clot or plaque buildup in the arteries supplying nutrients to the brain tissue. There are two main types of stroke: ischemic, which accounts for more than 85% of the cases, and hemorrhagic. According to Tsao et al. (2022) and the World Health Organization, 15 million people worldwide suffer a stroke every year, and its prevalence increases with age. Large Vessel Occlusions (LVO) is a specific type of AIS and refers to the complete or partial blockage of one of the brain's major arteries. LVO, which encompasses both anterior and posterior circulation, is responsible for approximately 46% of acute ischemic strokes. Among these cases, around two-thirds of LVOs occur in the anterior circulation, primarily affecting the Internal Carotid Artery (ICA) and the Middle Cerebral Artery (M1). The remaining proportion occurs in the posterior circulation, considering the Vertebral Artery, Basilar Artery, and Posterior Cerebral Artery (PCA) — (see Fig.1a). Another type of occlusion that can occur is called Tandem, which occurs in less than 10% of the cases (Sweid et al., 2020) and refers to an occlusion in

more than one artery: a large blood vessel, such as ICA, and in an intracranial artery. The damage provoked by LVOs depends mainly on the location of the occlusion and on the time of blocking.

According to Martins-Filho et al. (2019), patients suffering from AIS related to an LVO experience the highest levels of morbidity and mortality, along with the lowest probability of achieving arterial recanalization through a clot-disolving medication, called intravenous thrombolysis. Nonetheless, recent trials have presented strong evidence supporting the effectiveness of endovascular mechanical thrombectomy in treating such cases. Consequently, the timely identification and transfer of patients with LVO to stroke centers have become imperative for facilitating prompt detection and providing appropriate endovascular treatment. Given the time-critical nature of mechanical thrombectomy, there is a need for efficient vascular imaging methods that can swiftly diagnose LVO (Mayer et al., 2020).

There are several imaging modalities for stroke imaging, being the most primarly Non Contrast Computed Tomography (NCCT), Computed Tomography Angiography (CTA) and Computed Tomography Perfusion (CTP). NCCT involves taking X-ray images of the brain without contrast agents in a way that provides information about the presence of bleeding, tumors, or other abnormalities. This is commonly used as the initial imaging modality for stroke patients since it helps rule out conditions that can mimic a stroke. Nonetheless, NCCT is not very sensitive to detecting early signs of stroke or small infarctions. Then, CTA is used to visualize the blood vessels in the brain. This imaging modality involves the injection of contrast dye into the patient, which helps to highlight the blood vessels (see Fig.1c). It is a reliable method to detect LVOs and it has been used to determine if a patient is a good candidate for a mechanical thrombectomy (Shafaat and Sotoudeh, 2022). Moreover, CTP uses a series of rapid CT scans, and is primarily used to evaluate the passage of blood through the tissues. It helps clinicians evaluate the tissue viability and identify areas of reduced blood flow.

As previously stated, endovascular thrombectomy (EVT) is the established treatment for patients exhibiting stroke symptoms within a 24-hour window, as its effectiveness diminishes beyond this timeframe. The value of time becomes increasingly crucial in such cases. The main goal of the treatment is to reestablish the blood flow as soon as possible, reducing the risk of permanent damage, improving outcomes after the episode, and minimizing the impact on the patient's neurological function. The optimal time window of this treatment is considered to be within 6 hours after the onset of symptoms. According to Sweid et al. (2020), every 30-minute delay decreases the odds for a favourable outcome by 11%. This is the main reason there have been improvements in several factors to reduce stroke care timing, including stroke assessment tools. Despite the efforts, there is still the need to standardize stroke detection and triage, which is time-sensitive (Murray et al., 2020). To streamline this process, the implementation of automated imaging-based tools for detecting LVO has demonstrated improvements in the timing of EVT decision-making, ultimately resulting in enhanced clinical outcomes. Although some commercial solutions have addressed this application, their performance and utility are difficult to compare since testing on a common dataset has not been performed. That is one of the main reasons why, this year, the challenge IACTA-EST 2023 [1], which addresses the use of CTAs for EVT stroke treatment, is proposed as part of the 2023 IEEE International Symposium on Biomedical Imaging (ISBI).

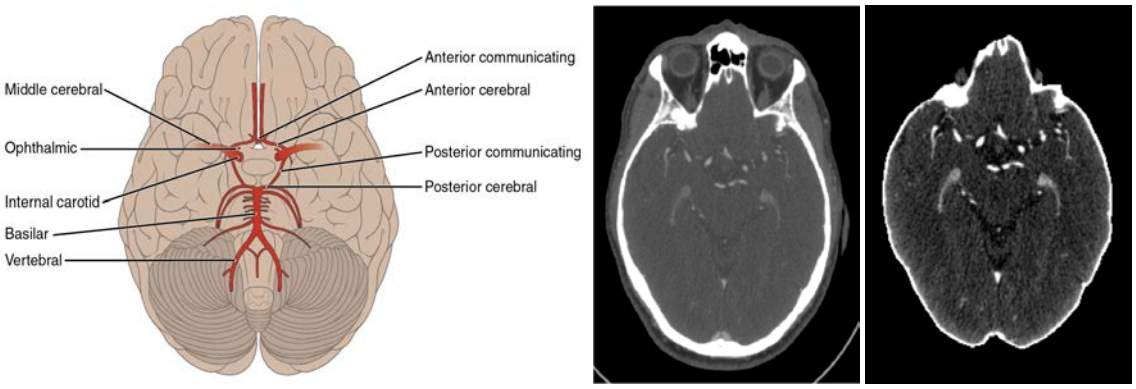
The IACTA-EST challenge deals with several important and difficult tasks regarding AIS, which are to provide a curated imaging dataset of brain CTAs from multiple clinical sites with evaluation metrics, to determine the presence or absence of an LVO in CTAs (task 1 of the challenge), and to obtain the brain reperfusion prediction using CTAs and clinical variables (task 2 of the challenge).

1.1. Our work

This master thesis is related to the first task of this challenge, which is to identify the presence of LVO in CTA images. However, we also want to determine the exact localization of an LVO based on CTA images. For this purpose, we will work with two different tasks, as shown in Fig. 2: binary classification of CTA images to determine if a patient has or does not have an LVO, and LVO detection, which consists of determining the 3D localization of a blood clot with a bounding box. For the first task, we propose conducting a comparative analysis between base deep learning networks employed for classification purposes, as opposed to incorporating brain symmetry as an input into the networks using two distinct strategies. The outcomes of this task will be assessed using two datasets: the IACTA-EST 2023 challenge dataset and a dataset obtained from Hospital Dr. Josep Trueta. The latter dataset was meticulously processed and annotated by the authors under the hospital's neurologist's guidance and validation. For the second task, we propose using a Retina UNet within the framework of nnDetection Baumgartner et al. (2021), which, to the best of our knowledge, has not been used for this purpose before. The approaches in this work for both tasks show some promising results.

The next sections are organized as follows: Section 2 introduces the current work around these two tasks regarding the classification and detection of an LVO. In Section 3, we provide information about the databases

[1] https://lgiancauth.github.io/iacta-est-2023/



(a) (b) (c)

Figure 1: Brain arterial circulation. (a) Arterial cerebral circulation. Taken from JoeNiekroFoundation (2017). (b) Axial view - CTA of the brain, (c) Brain extracted from original CTA and intensity processed.

used in this work and the different approaches used, including the structure of the experiments and the methods. The results obtained and the most relevant experiments are shown in Section 4. Finally, we will discuss and analyze the obtained results in Section 5, and give conclusions and future work in Section 6.

2. State of the art

In recent years, significant advancements have been achieved in the application of artificial intelligence (AI) within the medical field. A recent survey article by Chavva et al. (2022) highlights the development of diverse AI systems catering to various aspects concerning ischemic strokes. These systems encompass tasks such as distinguishing strokes from mimics, detecting large vessel occlusions (LVOs), evaluating the extent of reversibility in ischemic injuries, and aiding in diagnostic decision-making for selecting optimal candidates for endovascular therapy (EVT). Furthermore, several clinically validated software platforms have been specifically designed for these purposes. Notably, Viz.ai, iSchevaView RapidAI, Brainomix AI, MethinksLVO, and StrokeViewer are among the noteworthy examples mentioned by Murray et al. (2020) and Chavva et al. (2022), demonstrating their utilization in the identification of LVOs, diagnosis of ischemic or hemorrhagic strokes, and the assessment of potentially salvageable tissue.

The main issue with the existing software is that it tends to have problems integrating into a different data stream. Also, there is a lack of standardization to validate the systems, which makes it difficult to compare with new algorithms (Chavva et al., 2022). Although AI can be used in clinical practice for stroke management in several ways, in this section, we will focus mainly on two tasks: classification and detection of thrombi causing the LVO (see Fig.2).

2.1. Classification

Over the last years, the use of deep learning for thrombus finding and classification has been increasing. An example of an article that uses CTAs for LVO is Stib et al. (2020) which proposes a 2D approach to classify the presence or absence of an LVO, taking only the slices from the skull vertex through the circle of Willis (see Fig. 1a). They use the CT angiographies' three phases (arterial, peak venous, and late venous phases) to experiment with different combinations. They obtained an AUC of 0.89, sensitivity of 100% and specificity of 77% using the combination of the three phases. There is also the use of 4D-CTA for detecting occlusions in the intracranial anterior circulation, presented by Meijs et al. (2020). This approach is able to detect patients that are eligible for endovascular therapy with high numbers of sensitivity, specificity and AUC: 95%, 92% and 0.98 respectively. However, this method does not provide a direct localization of the occlusion. Moreover, it uses 4D-CTA, which is less commonly available in hospitals than CTA.

In the same manner, Barman et al. (2019) presented the first publication outlining the algorithm and methodology for utilizing CTA images in the detection of AIS. The authors propose DeepSymNet, a convolutional neural network (CNN) that leverages the brain's symmetry to address image classification. Their approach involves working with 3D representations of the brain's hemispheres and incorporating inception modules to facilitate the network's ability to discern differences between them. Building upon this work, Czap et al. (2022) introduced an enhanced version of the same algorithm by adding symmetrical and nonsymmetrical pathways. DeepSymNet is currently on its 3rd version, developed by Giancardo et al. (2023), in which they input two hemispheres separately into a network composed by 3D VGG blocks, with shared weights between the two data paths. The primary objective of their research is to generate a full segmentation of the stroke core using the

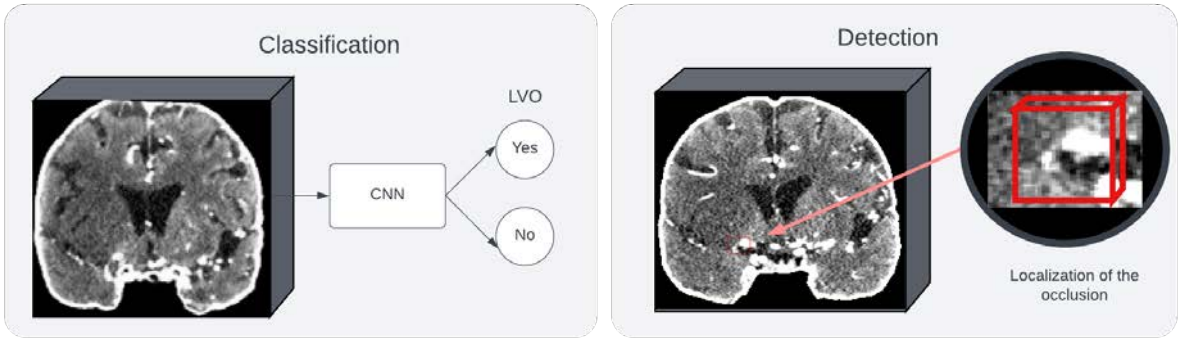


Figure 2: Differences between the two tasks in this study.

deep learning architecture.

There have also been some articles that use a specific software from the ones mentioned before to validate their results and their use in different clinical data. An example is Luijten et al. (2022), which uses the Nico.lab algorithm StrokeViewer[2] to obtain a binary input considering if an LVO is detected or not. Their results are based on detecting the right position of the clot found, and not based on the binary classification. However, they showed lower values of sensitivity and specificity for M2 occlusions. In the same way, Yahav-Dovrat et al. (2021) used Viz LVO[3] in real clinical practice at a comprehensive stroke center to determine if a patient has an LVO, achieving sensitivity of 82%, which is similar to other studies with the same purposes working with different versions of Viz LVO software, such as Barreira et al. (2018) and Chatterjee et al. (2019).

2.2. Detection

Based on the articles reviewed for this work, most of them use already-developed software to have the localization of the thrombus. However, there are few articles with information about details and explanation of how this localization is performed. In other cases, researchers use a region of interest or bounding box provided by experts, using NCCT images or CTAs. This could be mainly because detection tasks are commonly associated with the segmentation of the thrombus causing an LVO. Therefore, the final goal of most of the work done around this problem is to do a segmentation of the probable blood clot or thrombus.

One article that deal with segmentation using a previously obtained ROI, is Lucas et al. (2019). In this work, the authors compute the ROI on the training data including all the MCA and ICA clot segmentations plus a margin of 5 voxels in each direction. Then, they use this ROI as input of a Unet to obtain the segmentation.

The work proposed by Tolhuisen et al. (2020) is one of the articles which explains the process of finding the localization of an LVO on the anterior circulation system. They propose two patch-based CNNs based on AlexNet. The first one focuses on the asymmetry between the two hemispheres of the brain, and the second network focuses on the detection of Hyperdense Artery Sign (HAS), which is one of the earliest signs of ischemic stroke, and its attenuation is correlated with the concentration of red blood cells in the thrombus. If a patch was classified as having a thrombus, they performed a voxel-wise segmentation for the specific patch. They obtained promising results, but the volumetric and spatial agreement of their findings is low. In this approach, it is also necessary to register all the images - CTAs and NCCTs- to be able to work with symmetrical differences.

In the study proposed by Mojtahedi et al. (2022), they use the StrokeViewer LVO software from Nico.lab to create a bounding box around the probable location of the LVO. The system uses both NCCT and CTA images. After this, they use a dual-modality U-Net for segmentation. There are some cases in which the software used is not able to find LVO, most of them being M2 type of occlusions. Another study that uses the same software for LVO localization is presented by Bruggeman et al. (2022), in which the main goal is to test the algorithm given by the vendor in a different clinical dataset. Nico.lab performed the training and testing with more than 1000 CTAs. Therefore, detailed information about it is not publicly available, but they mention they register the images to the MNI space and find a bounding box mask, which is considered correct if it's on the right hemisphere, and covers part of the exact occlusion. They mainly show problems and the most distal occlusions. They also show a low rate of false positives detailing reasons why the system may fail.

The purpose of all these investigations is to identify occlusions within the anterior circulation system, with limited emphasis on occlusions in the posterior circu-

[2]https://www.nicolab.com/strokeviewer/
[3]https://www.viz.ai/lvo-ctp

lation. Primarily, the studies predominantly employed a combination of non-contrast CT (NCCT) and CTA imaging or solely relied on NCCT. To our knowledge, scientific literature lacks descriptions of models for automatic occlusion detection using only CTA images for anterior and posterior circulation. It is worth noting that previously mentioned software packages represent exceptions to this observation. Nonetheless, those encompass undisclosed algorithms.

In the year 2021, Baumgartner et al. (2021) proposed a self-supervising method for medical object detection, called nnDetection, that adapts itself to arbitrary medical detection problems, also acting as a standard interface for different data sets. Its effectiveness has been demonstrated on other medical imaging tasks. For instance, challenges such as LUng Nodule Analysis (Luna) and Aneurysim Detection And segMentation Challenge (ADAM), which uses Time of Flight Magnetic Resonance Angiongraphies. They also provide information guides for the detection task on different organs such as Liver, Pancreas, Prostate, Lung, Colon, etc.

3. Material and methods

3.1. Data

3.1.1. IACTA-EST Dataset

The dataset used for the classification part of this project is from the Image Analysis for CTA Endovascular Stroke Therapy (IACTA-EST) Challenge, which aims to provide a curated image dataset from multiple clinical sites in order to minimize the gap between current studies and commercial solutions not being able to compare results with evaluation metrics. All of the data used is from the first task of the challenge, phase 1, and contains images after a preprocessing pipeline that involves: conversion from DICOM to Nifti format, resampling and registration to a common image space with rigid transformation, skull stripping, and intensity values clipped to a range from 0 to 100 Hounsfield Units (HU). Only LVOs from the anterior circulation system are considered in the dataset, being them occlusions in the ICA, M1, M2 or A1 brain vasculature. The dataset contains 301 cases, from which 142 were classified as having an LVO and 159 as not having what is considered an LVO in the challenge. For more details about image resolution, voxel spacing and slice thickness, refer to Table 1.

3.1.2. ICTUS Dataset - Trueta Hospital

For the detection part of this project, we worked with a dataset acquired in the Hospital Dr. Josep Trueta, Girona, Spain. It consists of 321 cases with a valid CTA, with all the patients having an LVO. The localization of the occlusions was distributed as follows: 46% M1, 18% M2, 13% ICA, 7% Tandem, 6% Basilar, 5% PCA occlusions, and 5% corresponding to M3, A2, VA, and extracranial occlusions. After discarding the extracranial occlusions, and cases where there was uncertainty about the exact localization of the LVO, we were left with a total of 310 valid cases. All of the examinations were performed with a Philips Healthcare Ingenuity CT scanner. More information about the imaging protocol can be found in Table 1. All the valid cases will be used for inference in the LVO classification as either present or absent. Conversely, only 124 cases will be used for the detection task due to the availability of ground truth thrombus segmentation obtained by manual annotations using ITK-snap. These annotations were carried out by the authors of this master thesis along with the guidance and validation of an expert neurologist from the Hospital. The annotations were not done in all of the slices of the CTA cases but just in the main ones where the size and shape of the thrombus were appreciated.

Table 1: Image characteristics of the different datasets.

Dataset	**Resolution**	**Voxel Spacing**	**Slice Thickness**
IACTA-EST	146x182x133	1x1x1	Unknown
Trueta Hospital Classification	146x182x133	1x1x1	0.9
Trueta Hospital Detection	512x512x378	Variable	0.9

3.2. Data pre-processing

For preprocessing the hospital dataset, we performed different steps according to the task, as graphically shown in Fig. 3.

- Classification: For the hospital dataset to be used as an inference set for the classification experiments, we followed the preprocessing detailed in Giancardo et al. (2023), which has the same steps as the images of the IACTA-EST challenge used for training, explained in 3.1.1. The final input of the network consisted of the image after registration, cutting of blank spaces, and clipping the brain intensities between 0-100 HU.
- Detection: For the detection task, we performed the conversion from DICOM to Nifti format, then we used **fsl** (Jenkinson et al., 2012) for the preprocessing. First, we used *robustfov* to focus on the skull in the images, given that the original images show not only the head but also the patient's upper body. After this, we used *bet* for skull stripping. In the end, we obtained just the brain in the CTA images. For the next step, we clipped the intensity values of the brain between 0-200 HU according to the hospital doctors' suggestions. We did not register or resize the images to use them as input for the nnDetection framework.

3.3. Methodology

This study is organized into two primary stages. First, the IACTA-EST dataset from phase 1, task 1, will be

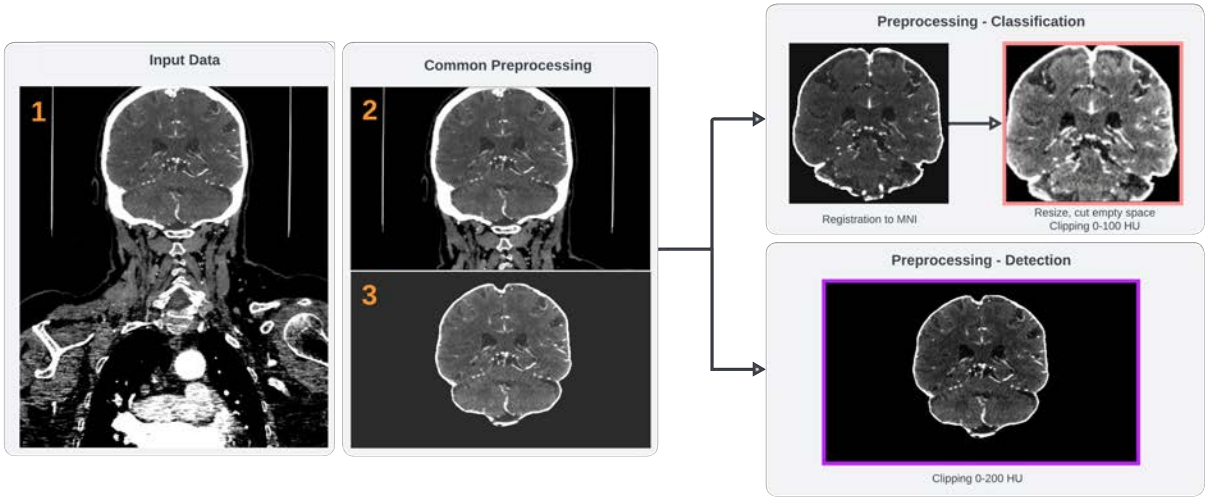


Figure 3: Preprocessing Pipeline. 1. Image in nifti format. 2. Result after using *robustfov*. 3. Result of using *bet*. The next steps change according to the task. The images highlighted are the final output for each problem. For detection we only perform clipping of intensities, although for classification, we also register and crop the image.

utilized to train models employing various strategies to identify the presence of an LVO in CTA images. Additionally, the hospital dataset will be employed for inference during this classification phase. The second stage involves training a detection model using the nnDetection framework on the hospital dataset to precisely localize the occlusion within the CTA images.

Furthermore, as part of the comparison process, a baseline DeepSymNet-V3 classifier proposed by Giancardo et al. (2023) will be trained for binary classification. This will allow for a comparison of results with the proposed networks, which are based on the same principle and will be further elucidated below.

In the next paragraphs, we will describe each stage with the methods proposed, the training process of the models, and evaluation metrics used to assess their performance.

3.3.1. Classification

For the classification component, we will conduct a comparative analysis between two approaches to determine the presence or absence of LVO in a CTA scan. The first approach employs a non-symmetry strategy, while the second approach involves utilizing the symmetry information from the two brain hemispheres within the network.

Non-Symmetry approach.

For the non-symmetrical path, we will use the 3D CTA volumes as input of different CNNs, obtaining a binary output as LVO present or absent. This is graphycally explained in Fig. 4.

Symmetry approach.

Regarding the symmetry path, we will explore two strategies:

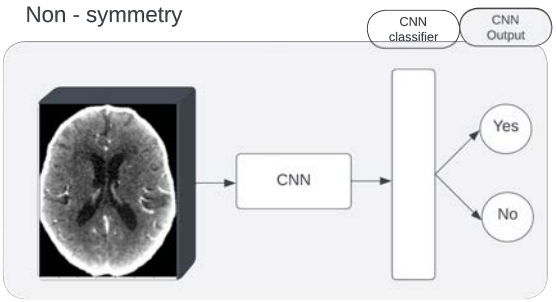


Figure 4: Non-symmetry approach: 3D CTA is the input of a CNN network, which will return the output of having or not an LVO.

1. Use a symmetry approach inspired by Barman et al. (2019) with DeepSymNet and its subsequent versions, in which, to compare the disparities between the two brain hemispheres, we employ two branches having identical architecture, each dedicated to processing a distinct hemisphere. To capture the disparities between the outputs of these two branches, we experiment with different approaches that are summarized in Fig. 5 and described below:
    - Experimenting with the network used in each branch sharing weights between the two paths against having an independent set of weights (represented in Fig. 5b)
    - Using the difference module L-1, as introduced in Barman et al. (2019), which takes the absolute difference between the high-level convolution filter outputs corresponding to each hemisphere. This L-1 module serves as the input of the CNN classifier, whose

output will go through a Multi-layer Perceptron (MLP) that will be used to learn the information contained within it, considering it has crucial information about differences between the brain hemispheres. See Fig. 5c, option A.

- Concatenating the outputs of each branch, representing the different hemispheres, and using this concatenation as the input to an MLP, which learns the differences between the hemispheres. See Fig. 5c, option B.

2. Use a symmetry approach that involves stacking both hemispheres of the brain and feeding them into a single network. Additionally, an MLP is incorporated before producing the final output. Refer to Fig. 6 for a clearer depiction of this strategy.

Architectures.

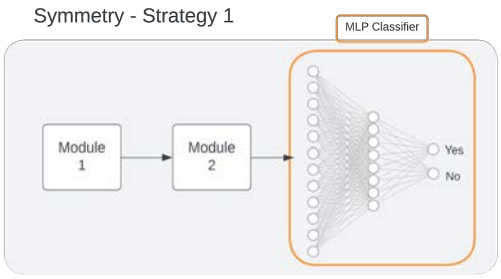
For this study, we compared the performance of three different networks:

- Resnet-18: It is a convolutional neural network CNN introduced for the first time in 2015 (He et al., 2015). The architecture of this network is shown in Fig. 7. We worked with three different initialization of weights for this network:

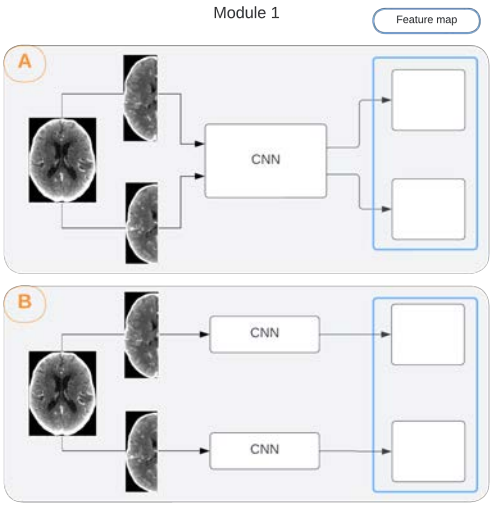
  1. Resnet-18 trained from scratch, with random weight initialization.
  2. Resnet-18 pretrained with videos, considering spatial and temporal components, and 3 input channels (RGB) - (Tran et al., 2018). Since this network was pretrained with RGB videos, and we have grayscale images, we replicated the same input for the three channels.
  3. Resnet-18 pretrained with medical images, obtained from Chen et al. (2019). The weights released were trained with data from different organs such as brain, prostate, liver, heart, pancreas, etc. This network was pretrained with grayscale images, so we did not perform any adaptation on the input data.

- Densenet121: It is a CNN introduced by Huang et al. (2017). The architecture of the network is shown in Fig. 8.

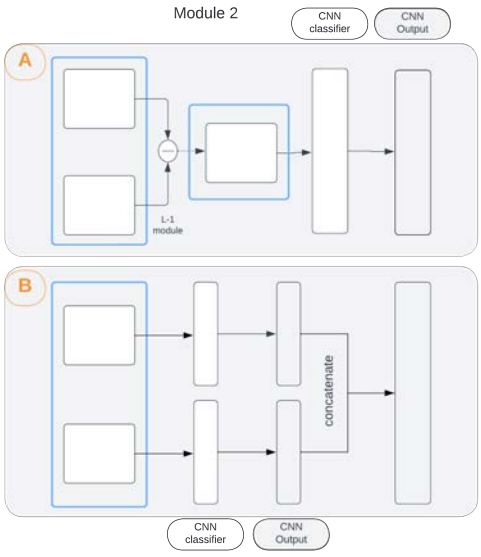
- DeepSymNet-V3: 3D CNN proposed by Giancardo et al. (2023), which receives as input two images, each one having it's own path. The model it's composed by three 3D VGG blocks, sharing weights between the two paths. These two paths are combined by an L1-layer, performing the difference between the feature map of each branch, but still preserving the spatial information.



(a) Symmetry approach - Strategy 1: Composed by Module 1, that would determine if each hemisphere training path would share or not weights between them. This module produces an output (two feature maps) that feeds into Module 2. Module 2 will determine how to combine and obtain the difference between the two hemisphere training paths. This module will return the results of the CNN classifier with the original number of classes per model. The output from Module 2 serves as an input for an MLP Classifier.



(b) Module 1 - Strategy 1: The initial CTA brain volume is divided by hemispheres. Each hemisphere would be the input of a CNN model, having independent training paths. In option (A), the CNN models corresponding to each hemisphere share weights between them. In option (B) each CNN model has independent weights initialized in the same way .



(c) Module 2 - Strategy 1: Given two feature maps, corresponding to each hemisphere of the brain respectively, option (A) represents the use of L-1 module, obtaining a feature map with the absolute difference between them, that will then pass through the CNN classifier. Option (B) represents each path passing directly through the CNN classifier obtaining an output corresponding to each branch. For combining the two outputs, in this option, we concatenate them.

Figure 5: Explanation of Symmetry Strategy 1.

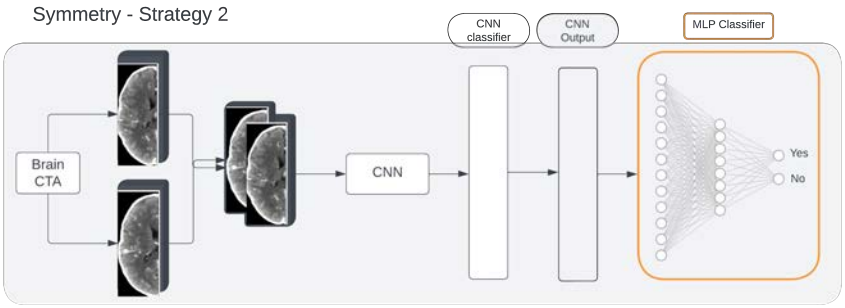


Figure 6: Symmetry approach - Strategy 2: The initial CTA brain volume is divided by hemispheres. Both hemispheres are stacked together, creating an input of two channels that will be fed into the CNN network. This network will output its original number of output neurons, which will next be the input of an MLP classifier.

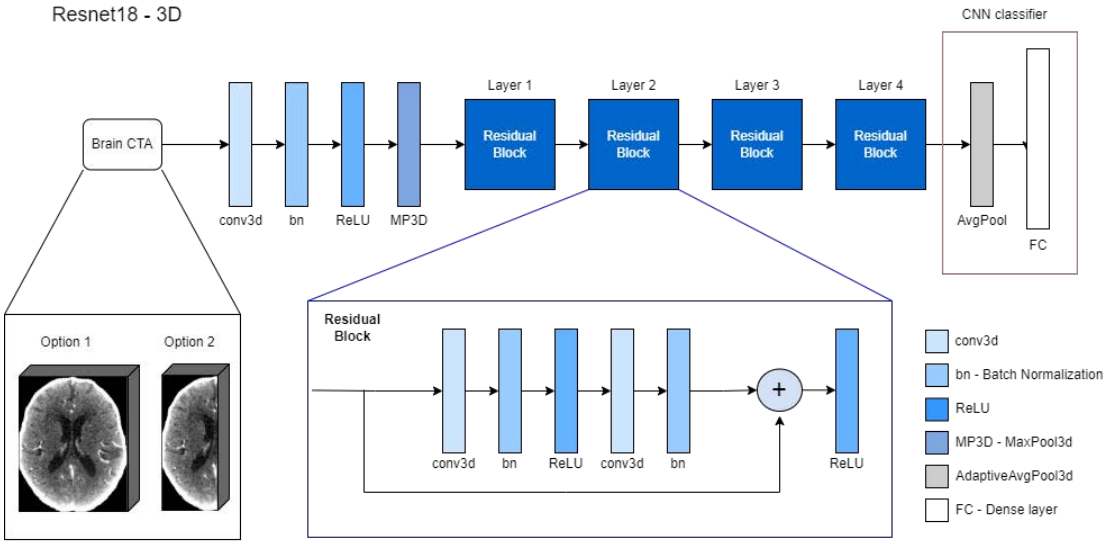


Figure 7: Resnet-18 architecture used in this study. The input will be either the full brain, or the right or left right-flipped hemisphere. All the weights belonging to the blocks in different shades of blue can be frozen. The CNN classifier that is remarked in the image changes according to what we need, either to obtain a final prediction, or to obtain what will be the input of an MLP Classifier.

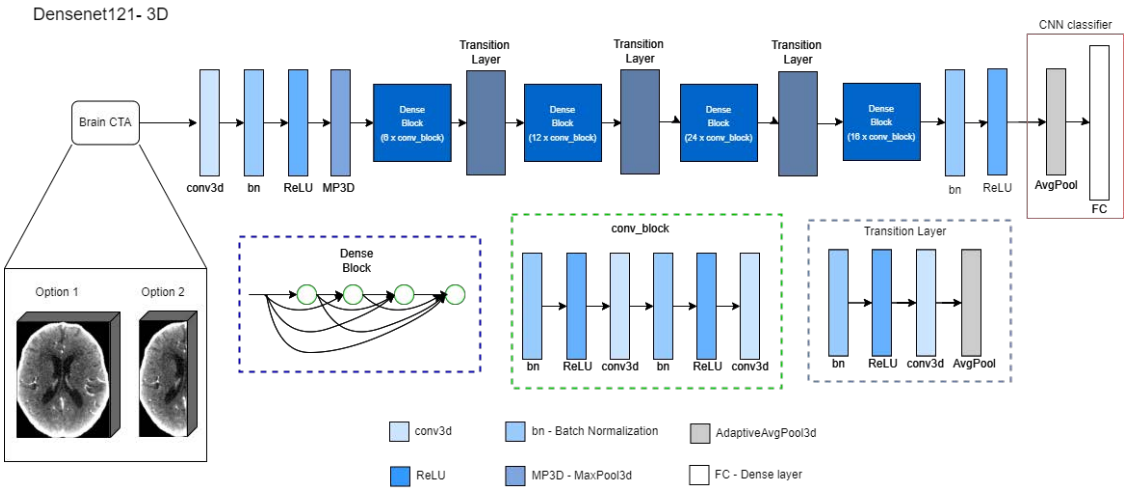


Figure 8: Densenet-121 architecture used in this study. The input will be either the full brain, or the right or left right-flipped hemisphere. All the weights belonging to the blocks in different shades of blue can be frozen. The CNN classifier that is remarked in the image changes according to what we need, either to obtain a final prediction, or to obtain what will be the input of an MLP Classifier.

The CNNs used in this approach have their original building blocks. Originally, Resnet-18 has 400 output neurons, and Densenet-121 has 1000. In some cases, we will use the original classifier and in others, we will make changes in the CNNs classifier according to our needs. Regarding the symmetry approach, we will add an MLP classifier at the end of the different combinations to help the network to learn the differences between the two brain hemispheres. The MLP input layer will adapt to the output neurons of the CNNs. The input layer determines the MLP number of hidden layers and neurons as shown in Table 2.

Table 2: MLP configuration for symmetry approach, according to the difference strategy selected.

Model	Difference strategy	Input	Hidden layers	Output
Resnet18	L-1 module	512	256	2
Resnet18	concatenation	800	512, 256	2
Densenet121	L-1 module	1000	512, 256	2
Densenet121	concatenation	2000	1000, 512, 256	2

Training plan.

For this classification task, the dataset from the challenge IACTA-EST described in 3.1.1 will be used for training. From the available 301 cases, 15% is used for testing following a balanced positive and negative label distribution. The remaining 255 cases are used for train and validation, considering 5-fold cross-validation experiments. The positive/negative label distribution is shown in Table 3.

In the initial phase of the experiments, we began by exploring models following a non-symmetrical path. This approach involved utilizing a single path for each network, with the input being the entire brain comprising both hemispheres. We conducted parameter tuning, testing two optimizers (adam and stochastic gradient descent - sdg) and experimenting with different learning rates (LR) ranging from 0.01 to 1e-5. The optimal LR varied depending on the specific model employed; however, adam optimizer consistently yielded the best results. Additionally, we analyzed the impact of freezing different layers in the pretrained networks to identify the optimal combination of frozen layers for our data. The findings from this initial phase included the best parameters (LR and optimizer) and the most effective combination of frozen layers for the pretrained models. These outcomes will be used for the next phase of the experiments.

In the subsequent set of experiments, we built upon the insights gained from the previous phase - specifically, the optimal LR and frozen layers combination. For each network architecture mentioned in section 3.3.1, we explored two sharing weight strategies and two approaches for obtaining and comparing the differences between the paths of each brain hemisphere. With the first symmetry strategy, we considered four distinct combinations for each architecture configuration. For the second symmetry strategy, we leveraged the successful outcomes of the initial phase, modifying the architectures to accommodate 2 input channels instead of 1: we stack the two brain hemispheres instead of each one going through a different path to then compare and combine the results. The 2 input channel adaptation was possible for all cases except for the Resnet-18 pretrained with videos as we had to give unwanted importance to just one hemisphere of the brain by replicating it on the third channel. Therefore, we decided the following configuration for this network: channel 1 - left hemisphere, channel 2 - left-right flipped hemisphere, channel 3 - the absolute difference between both hemispheres.

Table 3: Positive/negative label distribution from IACTA-EST challenge data set.

Labels	**Train**	**Validation**	**Test**	**Total**
0	114	21	24	159
1	102	18	22	142
Total	216	39	46	301

In the final step, we selected the most promising combinations from the experiments and employed the Trueta hospital dataset for inference. Within this dataset, all valid cases were considered as LVO present. To generate the results from this dataset, we employed majority voting based on the models trained using 5-fold cross-validation. Given that the hospital dataset encompasses occlusions in both the anterior and posterior circulation, whereas the training set (from the IACTA-EST dataset) only contained information about anterior circulation occlusions, we made the decision to conduct separate inferences for two categories: all occlusions in the dataset and anterior occlusions in the data. In the results section, we will present the performance of the models at each stage on the testing data (train-val-test split), as well as their performance on the aforementioned hospital dataset.

Data Augmentation and network input.

All of the models were trained using the same data augmentation techniques, consisting of:

- Random flipping ($\rho = 0.5$) around the x-axis, corresponding to the interhemispheric fissure of the brain.
- Normalization using $\mu = 0.485$ and $\sigma = 0.229$. This step was performed for all the images.
- Randomly chosen motion blur, median blur, or Gaussian blur, applied with $\rho = 0.6$.

For validation and testing data, only normalization was performed. It is worth mentioning that, for the symmetry approaches we first performed the data augmentation and then, we separated both hemispheres, right-flipping the left hemisphere of the brain, to use both brain hemispheres independently as input of the network, as shown

in Fig. 3. For non-symmetry approaches, we used the original dimension of the image.

Evaluation metrics.

For this task we will use the common metrics used in binary classification problems: Specificity, sensitivity, accuracy, and AUC-ROC. The specificity indicates the ability of the model to correctly identify healthy cases, in our case, LVO absent. On the other hand, sensitivity indicates the ability of the model to correctly identify diseased cases - LVO present. Accuracy represents the model's ability to correctly classify both positive and negative cases. Moreover, AUC-ROC is relevant since it allows the assessment of the model's performance across various thresholds for a comprehensive evaluation of the model's sensitivity and specificity trade-off. This metric is useful in medical problems to evaluate and optimize false positives and false negatives.

3.3.2. Detection

In the detection part, we use the framework nnDetection by Baumgartner et al. (2021), which has not been used before with CTAs, and also, not for this specific problem of localizing the LVO in a 3D volume. The instructions to adapt a new dataset are given in their GitHub repository [4]. We followed the instructions provided there and worked with the docker container of the framework. For this task, we manually split 10% of the available training cases to work as the test set. The framework internally creates a training plan with a 5 fold cross-validation system. The detection algorithm of nnDetection is based on Retina-UNet, although they are currently working on adapting other R-CNN detection networks. The framework expects the ground truth to be segmentation masks. Due to the complexity and intricacies of the LVO regions, obtaining precise 3D segmentations through manual annotations poses significant challenges. Despite our efforts, we encountered difficulties in achieving accurate and consistent 3D segmentations across all three views of our dataset. The process of manually delineating the LVO regions in three-dimensional space proved to be highly intricate, time-consuming, and prone to human error. As a result, we have opted to employ bounding boxes as an alternative representation for the LVO regions, since they provide a simplified yet informative approximation of the region of interest, encompassing the area where the LVO is most likely to occur. Therefore, we created bounding boxes around the manually annotated thrombus segmentation considering a margin of 2 pixels for all the cube coordinates. The final ground truth masks used in the framework consisted of 3D bounding boxes that enclosed the clot in the CTA image (see Fig. 9). All cases considered in this task had just one occlusion, and there were no cases without occlusion.

[4] https://github.com/MIC-DKFZ/nnDetection

Architecture.

The nnDetection framework uses a RetinaU-Net, proposed by Jaeger et al. (2020), combines the RetinaNet one-stage detector with the U-Net architecture, which is commonly used for semantic segmentation. This architecture complements object detection with semantic segmentation without introducing the additional complexity of previously proposed detectors.

Training plan.

This framework automatically creates a training plan according to the given dataset. It consists of several steps, including cropping, preprocessing of the input images, and details about the architecture and the input sizes of the images. For more information about the parameters optimization, refer to Fig. B.12. Given that the plan is specifically created based on the training data, and we did a manual split, we will mention some important considerations for the model training: the patch size used is $[128 \times 160 \times 128]$, and the target spacing is [0.45001221 0.58886719 0.58886719].

The model was trained with 124 cases, from which the 10% was used for testing, as explained in the data section 3.1.2. The data used for training and testing followed a similar distribution of occlusion localization, considering all the available cases.

The output of the nnDetection framework consists of a dictionary for each test case containing information about: predicted bounding boxes, prediction scores, predicted labels, original size of the raw data, origin of the image when read by itk, itk spacing and itk direction.

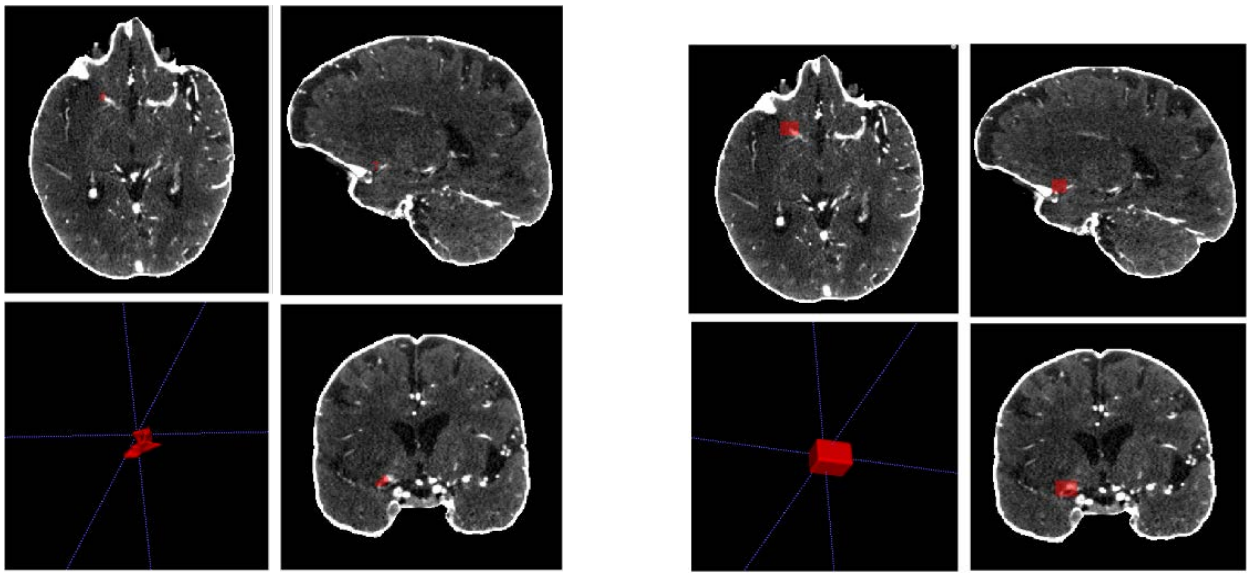
Data Augmentation.

For the data augmentation of this task, the configuration for transformations was the option 'BaseMoreAugmentation', with configurations such as scaling, rotation, changes in brightness, and additive noise. The configurations were used as default.

Evaluation metrics.

For this detection problem, we will consider Intersection over Union (IoU) and sensitivity, based on True Positives (TP), False Positives (FP), and False Negatives (FN). The IoU measures the overlap between the predicted and the ground truth regions of an object or region of interest (ROI). It ranges between 0 and 1, where a higher value indicates a better overlap. Given that we are working with medical data in 3D, we will consider a correct prediction with an IoU of 0.1 given that this value respects the clinical need for coarse localization and exploits the non-overlapping nature of objects in 3D according to Jaeger et al. (2020).

To obtain the number of TP, FP, and FN, we considered different confidence predictions and a minimum IoU of 0.1 with the ground truth to be considered as TP.



(a) Original label to CTA brain extracted image.

(b) Bounding-box to CTA brain extracted image.

Figure 9: Images and segmentation labels for detection problem.

Considering that all of the cases used for this task contained just 1 occlusion, we took the one with the greatest confidence score in case of finding more than one TP. For FP, we counted them as 1 if the bounding boxes showed a large intersection between them.

4. Results

In this section, the performance of the different strategies applied for both classification and detection tasks are presented.

4.1. Classification

We show the results of the different strategies applied for binary classification of the CTA images as LVO present or absent using the IACTA-EST 2023 dataset. All the results presented were obtained with 5-fold cross-validation models. As mentioned in the Section 3, we studied mainly two CNNs for approaching this task: Resnet-18 and Densenet-121. Considering Resnet-18, two different pretrained networks were compared. The first experiment performed consisted on comparing the network's performance based on freezing a different amount of layers to work with the ones that showed better results in the following experiments. We did this to consider computational resources used, time of training and the models' performance. The results for the Resnet-18 pretrained with MedicalNet weights, are shown in Table 4. The same experiment was performed for the Resnet-18 pretrained with videos, and the results show a similar tendency. The results of this experiment are added in the annex, in Table A.8.

Considering these findings, we opted to utilize both pretrained models (MedicalNet and 3d-videos), along with their two most successful combinations for the subsequent phase. Given that we had several experiments to try, we performed an extra selection between these two combinations, to choose the model that suits better our goals. Based on the results shown in Table 5, we could not appreciate a significant improvement in not freezing any layer, considering that the network consumes much more resources and takes longer to train. Therefore, we determined that the pre-trained models with the best trade-off between performance and computational resources and computational time were the ones frozen until the 2nd layer. Consequently, the following experiments will use both pretrained models with 2 layers frozen for comparison.

Table 4: Experiment of Resnet-18 pretrained with MedicalNet weights, showing the performance of freezing a different amount of layers. Highlighted results show the best sensitivity and consistent results in the other metrics.

Frozen layers	Accuracy	Sensitivity	Specificity	AUROC
All	0.587	0.227	0.917	0.619
3	0.674	0.455	0.875	0.672
2	**0.804**	**0.682**	**0.917**	**0.815**
1	0.783	0.546	1.000	0.942
None	**0.717**	**0.682**	**0.750**	**0.831**

Table 5: Comparison between Resnet-18 pretrained with videos (Resnet18-Videos) and pretrained with MedicalNet (Resnet18-Medical), with the best combination of frozen layers.

Model	Accuracy	Sensitivity	Specificity	AUROC
Resnet18-Videos-2	0.700	0.600	0.792	0.767
Resnet18-Videos-None	0.691	0.527	0.842	0.792
Resnet18-Medical-2	0.744	0.600	0.875	0.816
Resnet18-Medical-None	0.739	0.591	0.875	0.840

As part of this task, we want to compare the results

of exploiting symmetry through the network against not forcing the model to encounter differences between both hemispheres of the brain. For the non-symmetry path, we used a simple approach of classification using the networks mentioned before, in which the input image is the whole CTA of the brain. For the symmetrical path, we proposed two strategies: Srategy 1, as seen in Fig. 5, has 4 combinations to experiment on. The possible combinations are to share or not weights, and to use L-1 difference module or concatenate the outputs of each path. This strategy involves training two separate models, each dedicated to one hemisphere of the brain, and subsequently combining the obtained results. Strategy 2 instead, trains just one model, but using both hemispheres of the brain in separate channels. We also compared the results obtained with DeepSymNet-V3, the baseline from which we got the inspiration for the proposed strategies. The results of these experiments can be seen in Table 6.

From the Non-symmetry approach, we can see that Densenet-121 shows the best performance in all the metrics considered. In the symmetry approach - strategy 1, the results show a tendency for better performance of two models: Resnet18-Scratch and Resnet18-MedicalNet. This can be seen through the different metrics considered. The results of the symmetry approach - strategy 2 show that Resnet18-Videos obtains good and consistent performance between the metrics considered.

4.1.1. Inference on Trueta's hospital dataset

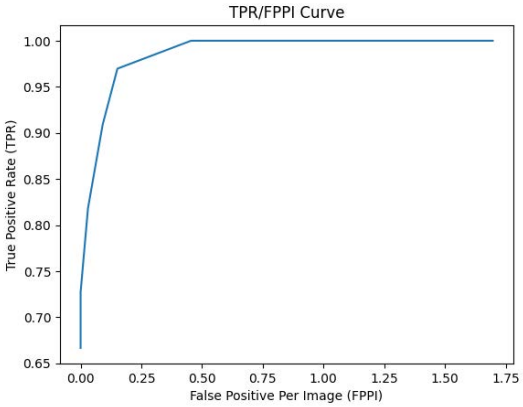
As part of our experiments, we tested the best performing models in the Trueta's hospital dataset, completely independent of the IACTA-EST challenge data used for training. The actual label for all the cases is LVO present, as the dataset did not have any case without LVO. Results of the inference process are shown on Table 7.

4.2. Detection

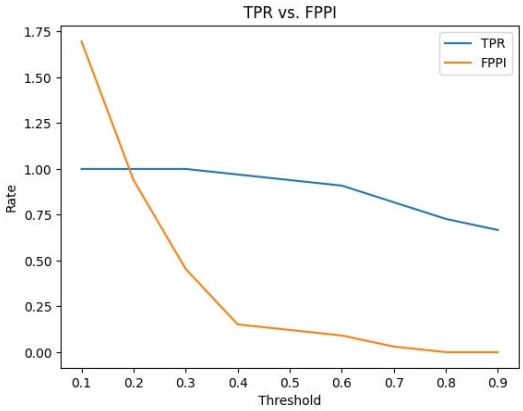
In this subsection, we present the results of nnDetection over 11 cases used just for testing. As mentioned in Section 3, we had 124 valid cases, from which 10% was used just for training. The remaining 90% was used for training and validation. We performed the experiments with 3 cross-validation folds.

The overall results of the nnDetection model considering the testing cases, are shown in Fig. 10. Based on Fig. 10b, we chose a threshold of 0.4 of confidence prediction score to maximize the number of TP and minimize the FP. With this thresholds, 2 folds were able to find the right localization of the occlusion in all of the cases, and 1 fold failed in 1, leveraging a general sensitivity of 97%. As for the FPs, 2 models found 1 FP in two images, and 1 model found 1 FP in one image, leading to an approximate 0.15 FPpI, as seen in Fig. 10a.

Some of the results obtained from the nnDetection model are shown in Fig. 11. From the images we



(a) TPR/FppI curve considering the mean results of 3 folds, through different thresholds of prediction scores.



(b) TPR and FppI rate through different confidence score thresholds.

Figure 10: Results of nnDetection considering the mean results of 3 cross-validation folds.

can appreciate that the bounding box generated can precisely detect the occlusions. One interesting case is presented in Fig. 11b, where the model encounters an occlusion that is not actually there. After visually inspecting the image, we can confirm that there is a difference in contrast following the arteries, that can be confused with an LVO.

5. Discussion

This study has two main goals: the binary classification of CTA images in LVO present/absent, and the detection considering the exact 3D localization of the LVO.

5.1. Classification

For the classification part, we propose a comparative analysis between deep learning networks using information about symmetry against not using it.

Existing literature supports the notion that incorporating symmetry information as input to the network leads to improved outcomes. This observation stems from the fact that, in most cases, an LVO primarily affects one

Table 6: Results of different approaches for binary classification of an LVO. Consider the following abbreviations: S = shared weights, nS = not shared weights, D = Use of difference module, nD = not using Difference module but concatenation.

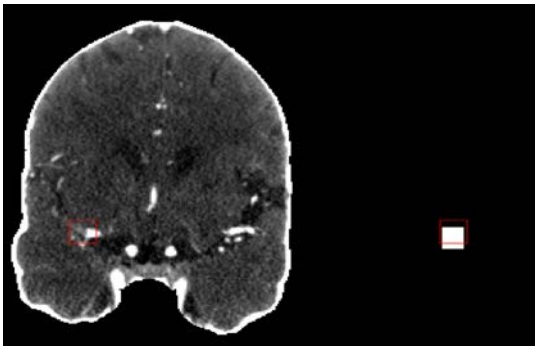
Model	Accuracy	Sensitivity	Specificity	AUROC
Non-symmetry				
Resnet18-Scratch	0.670	0.582	0.750	0.729
Resnet18-Videos	0.700	0.600	0.792	0.767
Resnet18-MedicalNet	0.743	0.600	0.875	0.816
Densenet121	**0.791**	**0.664**	**0.908**	**0.877**
Symmetry - strategy 1				
Resnet18-Scratch-S-D	**0.852**	**0.773**	**0.925**	**0.872**
Resnet18-Scratch-nS-D	**0.843**	**0.809**	**0.875**	**0.895**
Resnet18-Scratch-S-nD	0.726	0.718	0.733	0.789
Resnet18-Scratch-nS-nD	0.739	0.709	0.767	0.811
Resnet18-MedicalNet-S-D	0.822	0.745	0.892	0.864
Resnet18-MedicalNet-nS-D	0.774	0.664	0.875	0.813
Resnet18-MedicalNet-S-nD	**0.861**	**0.745**	**0.967**	**0.885**
Resnet18-MedicalNet-nS-nD	**0.857**	**0.764**	**0.942**	**0.899**
Resnet18-Videos-S-D	0.783	0.773	0.792	0.874
Resnet18-Videos-nS-D	0.787	0.782	0.792	0.870
Resnet18-Videos-S-nD	0.691	0.655	0.725	0.713
Resnet18-Videos-nS-nD	0.674	0.600	0.742	0.708
Densenet-121-S-D	0.800	0.736	0.858	0.843
Densenet-121-nS-D	0.752	0.618	0.875	0.788
Densenet-121-S-nD	0.743	0.555	0.917	0.813
Densenet-121-nS-nD	0.704	0.573	0.825	0.773
Symmetry - strategy 2				
Resnet18-Scratch	0.704	0.573	0.825	0.766
Resnet18-MedicalNet	0.709	0.582	0.825	0.783
Resnet18-Videos	**0.757**	**0.791**	**0.725**	**0.858**
Densenet-121	0.804	0.664	0.933	0.869
Baseline				
DeepSymNet-V3	**0.791**	**0.850**	**0.727**	**0.857**

Table 7: Accuracy results for Trueta Hospital Data. The results were obtain using the models described for inference, without any pretraining on this independent dataset.

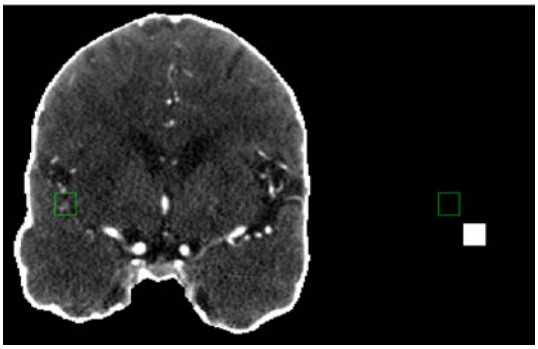
Model	Approach	All cases	Anterior circulation
Densenet121	Non-symmetry	0.5387	0.5909
Resnet18-Scratch-S-D	**Symmetry - strat1**	**0.6968**	**0.7727**
Resnet18-Scratch-nS-D	Symmetry - strat1	0.6677	0.7462
Resnet18-MedicalNet-S-nD	Symmetry - strat1	0.4548	0.5075
Resnet18-MedicalNet-nS-nD	Symmetry - strat1	0.6226	0.6970
Resnet18-Videos	Symmetry - strat2	0.6354	0.6931
DeepSymNet	Baseline	0.6419	0.7121

hemisphere of the brain, while the other hemisphere maintains normal blood circulation. Our experimental results reinforce this finding, as employing symmetry consistently enhances performance across all evaluated metrics. Notably, Resnet18-Scratch exhibits a significant approximately 10% enhancement in all categories following the integration of symmetry during training.

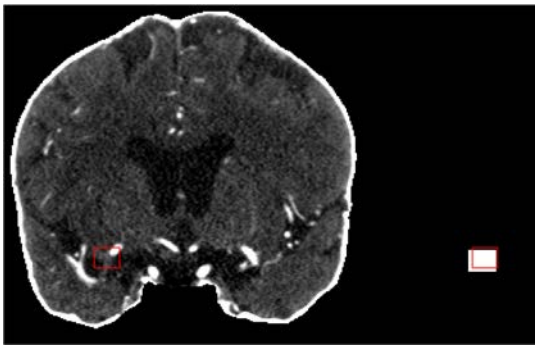
Regarding the use of brain symmetry, we proposed two distinct strategies and compared them. Both of the strategies came from the basic idea of dividing the brain into two hemispheres, with each hemisphere serving as an input for the network. The first strategy consisted of having each hemisphere of the brain following its own training path. From the results in the Table 6 in the first strategy compartment, comparing the different combinations proposed, we can see a general tendency of improvement when sharing weights between the two networks against not sharing them. Moreover, the numbers also show a tendency for improvement when using the difference module L-1 proposed by Barman et al. (2019). That is the case for most networks, except for the Resnet18 pretrained with MedicalNet, which was



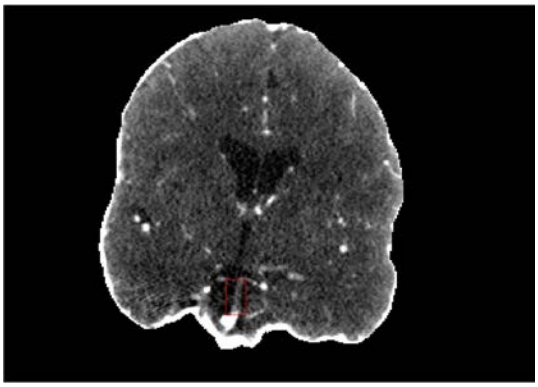
(a) True Positive detection.



(b) False Positive detection.



(c) True Positive detection on an M2 case.



(d) Inference over a posterior circulation image.

Figure 11: Bounding boxes results of nnDetection with a prediction score of 0.4. The white box located in the right, represents the ground truth of the occlusion.

used with two frozen layers. This case is particularly interesting because the feature map obtained before the L-1 module considers spatial information, and in the case of this pretrained network with frozen layers, the feature map at this point would probably not have enhanced differences between the two hemispheres. Therefore, the network works better with the second approach proposed, which is the concatenation of the outputs of each network.

The second strategy consisted of using both hemispheres of the brain as two input channels of the network. For these models, we modified the networks for them to accept the desired input size. Reviewing the results of Table 6 – Strategy 2 compartment, we can say that, in general, the first strategy showed better performance results. However, there is an interesting result regarding Resnet18 pretrained with Videos, which also had two layers frozen. This model shows the second best AUC, but the most consistent results between accuracy, sensitivity and specificity. We could attribute this to the difference in training for this network, which was originally pretrained for 3 input channels. Since each hemisphere corresponds to one input channel, we added the absolute difference between them in the third channel. This extra added information could have helped the model to learn more relevant features.

The highlighted models in Table 6 have similar performance results. It is worth mentioning that the emphasized experimets outperformed the DeepsymNet-V3 baseline for the testing portion of the IACTA-EST dataset. The model with the most consistent metrics and the second-best AUC is Resnet18-Scratch-nS-D. It is also relevant to mention that the pre-trained network did not significantly improve the general results.

From the best models, we used the combination of the 5-fold cross-validation results with majority voting to infer the binary classification of Trueta's hospital dataset. The results in Table 7 showed a 77% accuracy on correctly classifying CTAs with LVOs as LVO-present, using a completely independent data. These results encourage the seeking of information about the scanners used in the training dataset (IACTA-EST), and the feasibility of transfer learning for this problem.

5.2. Detection

For finding the exact 3D localization of the occlusion in CTAs, we used the nnDetection framework. The results shown in Fig. 10a show that the trained model is able to detect TPs with a very reasonable FppI rate. According to the results shown in Fig.10b, a threshold that optimizes the number of TP, minimizing the FP would be a confidence score of around 0.4.

Something worth analyzing is the localization of the occlusions in the cases used for testing. The data used for training this model was distributed in most of M1 cases, followed by M2, ICA, Tandem, Basilar and PCA. The occlusions in the test cases belonging to M2, ICA

and Tandem cases correctly detected by the 3 cross-validated models. For M1 cases, there exist the appearance of FP with a low general incidence. All of the occlusions in the test cases correspond to the Anterior Circulation system. Since some cases of Posterior Circulation were used in the training set, we used the model for inference of a Basilar occlusion case, from which we did not have the ground truth mask. However, by visual inspection we can prove that the model correctly detected the occlusion, as shown in Fig.11d. These promising results encourage the study to validate these results with more data, and with different type of occlusions.

Some of the drawbacks about using this model is the computational time it needs. For this dataset, of around 113 CTAs used for training and validation, each fold took around 6 days to train. Furthermore, the process of generating outputs for each image during prediction and inference typically requires approximately 5 minutes, not taking into account the additional time required for preprocessing and cropping that the framework performs automatically. In this study, we only performed 3 out of 5 folds, but for future work, we should perform a complete validation of all the cases.

6. Conclusions

This master thesis presents two main tasks: binary classification of CTA images in LVO present/absent, and detecting the clot in the image using the 3D localization. A comparative analysis is performed in the first task to determine the best combination of strategies and architectures to tackle the classification problem. We determined that the use of brain symmetry helps the network to determine the presence of an LVO in a CTA. Moreover, the L-1 module for combining the information of both hemispheres seems to help to have better results. However, depending on how the network is learning the fundamental features of the images, it may work better to combine them by simple approaches like output concatenation. The best combinations of model and symmetry approches gives an AUC of 0.87, 85% accuracy, 92% specificity and 77% sensitivity on test set. On the other side, this model also gives 77% accuracy on the independent Trueta hospital dataset.

For the second task, we propose the use of nnDetection for LVO detection. The model trained was able to find the right localization of the occlusion in all of the test cases, with a 0.15 FppI rate. This proposal shows very promising results that, with further experiments and clinical validation, could be used as a diagnostic tool for doctors to help with the time-sensitive LVO detection.

For future work, it is necessary to validate the classification models with more data, and also, to extrapolate their use for multi-class classification problems, which would give information about the brain branch in which the clot could be found. Also, it would be interesting to measure the impact of transfer learning for this problem, based on the results obtained on an independent dataset with models trained with information from multiple scanners. As for the detection task, it is necessary to clinically validate the results obtained, and to use nnDetection with a larger dataset, considering also less common occlusions for the model to be able to learn their features and find them.

Acknowledgments

I would like to express my gratitude towards my supervisors Dr. Xavier Lladó and Dr. Arnau Oliver for their guidance, encouragement, and support throughout this project. Also, I would like to thank the VICOROB research group for their help, especially, Uma Lal-Trehan for her contribution to this work, recommendations and help. Thanks also to Dr. Mikel Terceño for providing the dataset for this research, and for the patience and willingness to teach us and help us with the annotations. Thanks also to Dr. Luca Giancardo for providing the dataset information.

I would also like to thank the MaIA consortium and the friends I have made here. Last, but not least, I would like to thank my family, friends, and loved ones for their support and encouragement through these two years. I could not have done it without you.

References

Barman, A., Inam, M.E., Lee, S., Savitz, S., Sheth, S., Giancardo, L., 2019. Determining ischemic stroke from ct-angiography imaging using symmetry-sensitive convolutional networks, IEEE. pp. 1873–1877. doi:10.1109/ISBI.2019.8759475.

Barreira, C., Bouslama, M., Lim, J., Al-Bayati, A., Saleem, Y., Devlin, T., Haussen, D., Froehler, M., Grossberg, J., Baxter, B., et al., 2018. E-108 aladin study: automated large artery occlusion detection in stroke imaging study–a multicenter analysis.

Baumgartner, M., Jäger, P.F., Isensee, F., Maier-Hein, K.H., 2021. nndetection: a self-configuring method for medical object detection, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24, Springer. pp. 530–539.

Bruggeman, A.A., Koopman, M.S., Soomro, J., Small, J.E., Yoo, A.J., Marquering, H.A., Emmer, B.J., 2022. Automated detection and location specification of large vessel occlusion on computed tomography angiography in acute ischemic stroke. Stroke: Vascular and Interventional Neurology 2, e000158.

Chatterjee, A., Somayaji, N.R., Kabakis, I.M., 2019. Abstract wmp16: artificial intelligence detection of cerebrovascular large vessel occlusion-nine month, 650 patient evaluation of the diagnostic accuracy and performance of the viz. ai lvo algorithm. Stroke 50, AWMP16–AWMP16.

Chavva, I.R., Crawford, A.L., Mazurek, M.H., Yuen, M.M., Prabhat, A.M., Payabvash, S., Sze, G., Falcone, G.J., Matouk, C.C., de Havenon, A., Kim, J.A., Sharma, R., Schiff, S.J., Rosen, M.S., Kalpathy-Cramer, J., Gonzalez, J.E.I., Kimberly, W.T., Sheth, K.N., 2022. Deep learning applications for acute stroke management. Annals of Neurology 92, 574–587. doi:10.1002/ana.26435.

Chen, S., Ma, K., Zheng, Y., 2019. Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625 .

Czap, A.L., Bahr-Hosseini, M., Singh, N., Yamal, J.M., Nour, M., Parker, S., Kim, Y., Restrepo, L., Abdelkhaleq, R., Salazar-Marioni, S., Phan, K., Bowry, R., Rajan, S.S., Grotta, J.C., Saver, J.L., Giancardo, L., Sheth, S.A., 2022. Machine learning automated detection of large vessel occlusion from mobile stroke unit computed tomography angiography. Stroke 53, 1651–1656. doi:10.1161/STROKEAHA.121.036091. review citation 5.

Giancardo, L., Niktabe, A., Ocasio, L., Abdelkhaleq, R., Salazar-Marioni, S., Sheth, S.A., 2023. Segmentation of acute stroke infarct core using image-level labels on ct-angiography. NeuroImage: Clinical 37, 103362.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. arxiv 2015. arXiv preprint arXiv:1512.03385 14.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.

Jaeger, P.F., Kohl, S.A., Bickelhaupt, S., Isensee, F., Kuder, T.A., Schlemmer, H.P., Maier-Hein, K.H., 2020. Retina u-net: Embarrassingly simple exploitation of segmentation supervision for medical object detection, in: Machine Learning for Health Workshop, PMLR. pp. 171–183.

Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M., 2012. Fsl. Neuroimage 62, 782–790.

JoeNiekroFoundation, 2017. Brain basics.

Lucas, C., Schöttler, J.J., Kemmling, A., Aulmann, L.F., Heinrich, M.P., 2019. Automatic detection and segmentation of the acute vessel thrombus in cerebral ct, in: Bildverarbeitung für die Medizin 2019: Algorithmen–Systeme–Anwendungen. Proceedings des Workshops vom 17. bis 19. März 2019 in Lübeck, Springer. pp. 74–79.

Luijten, S.P.R., Wolff, L., Duvekot, M.H.C., van Doormaal, P.J., Moudrous, W., Kerkhoff, H., a Nijeholt, G.J.L., Bokkers, R.P.H., Yo, L.S.F., Hofmeijer, J., van Zwam, W.H., van Es, A.C.G.M., Dippel, D.W.J., Roozenbeek, B., van der Lugt, A., 2022. Diagnostic performance of an algorithm for automated large vessel occlusion detection on ct angiography. Journal of NeuroInterventional Surgery 14, 794–798. doi:10.1136/neurintsurg-2021-017842.

Martins-Filho, R.K.d.V., Dias, F.A., Alves, F.F., Camilo, M.R., Barreira, C.M., Libardi, M.C., Abud, D.G., Pontes-Neto, O.M., 2019. Large vessel occlusion score: a screening tool to detect large vessel occlusion in the acute stroke setting. Journal of Stroke and Cerebrovascular Diseases 28, 869–875.

Mayer, S.A., Viarasilpa, T., Panyavachiraporn, N., Brady, M., Scozzari, D., Van Harn, M., Miller, D., Katramados, A., Hefzy, H., Malik, S., et al., 2020. Cta-for-all: impact of emergency computed tomographic angiography for all patients with stroke presenting within 24 hours of onset. Stroke 51, 331–334.

Meijs, M., Meijer, F.J., Prokop, M., van Ginneken, B., Manniesing, R., 2020. Image-level detection of arterial occlusions in 4d-cta of acute stroke patients using deep learning. Medical Image Analysis 66, 101810. doi:10.1016/j.media.2020.101810. randomized control trials: Review this.

Mojtahedi, M., Kappelhof, M., Ponomareva, E., Tolhuisen, M., Jansen, I., Bruggeman, A.A.E., Dutra, B.G., Yo, L., LeCouffe, N., Hoving, J.W., van Voorst, H., Brouwer, J., Terreros, N.A., Konduri, P., Meijer, F.J.A., Appelman, A., Treurniet, K.M., Coutinho, J.M., Roos, Y., van Zwam, W., Dippel, D., Gavves, E., Emmer, B.J., Majoie, C., Marquering, H., 2022. Fully automated thrombus segmentation on ct images of patients with acute ischemic stroke. Diagnostics 12, 698. doi:10.3390/diagnostics12030698.

Murray, N.M., Unberath, M., Hager, G.D., Hui, F.K., 2020. Artificial intelligence to diagnose ischemic stroke and identify large vessel occlusions: a systematic review. Journal of NeuroInterventional Surgery 12, 156–164. doi:10.1136/neurintsurg-2019-015135.

Shafaat, O., Sotoudeh, H., 2022. Stroke imaging. StatPearls .

Stib, M.T., Vasquez, J., Dong, M.P., Kim, Y.H., Subzwari, S.S., Triedman, H.J., Wang, A., Wang, H.L.C., Yao, A.D., Jayaraman, M., Boxerman, J.L., Eickhoff, C., Cetintemel, U., Baird, G.L., McTaggart, R.A., 2020. Detecting large vessel occlusion at multiphase ct angiography by using a deep convolutional neural network. Radiology 297, 640–649. doi:10.1148/radiol.2020200334.

Sweid, A., Hammoud, B., Ramesh, S., Wong, D., Alexander, T.D., Weinberg, J.H., Deprince, M., Dougherty, J., Maamari, D.J.M., Tjoumakaris, S., et al., 2020. Acute ischaemic stroke interventions: large vessel occlusion and beyond. Stroke and Vascular Neurology 5.

Tolhuisen, M.L., Ponomareva, E., Boers, A.M., Jansen, I.G., Koopman, M.S., Sales Barros, R., Berkhemer, O.A., van Zwam, W.H., van der Lugt, A., Majoie, C.B., et al., 2020. A convolutional neural network for anterior intra-arterial thrombus detection and segmentation on non-contrast computed tomography of patients with acute ischemic stroke. Applied Sciences 10, 4861.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M., 2018. A closer look at spatiotemporal convolutions for action recognition, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 6450–6459.

Tsao, C.W., Aday, A.W., Almarzooq, Z.I., Alonso, A., Beaton, A.Z., Bittencourt, M.S., Boehme, A.K., Buxton, A.E., Carson, A.P., Commodore-Mensah, Y., et al., 2022. Heart disease and stroke statistics—2022 update: a report from the american heart association. Circulation 145, e153–e639.

Yahav-Dovrat, A., Saban, M., Merhav, G., Lankri, I., Abergel, E., Eran, A., Tanne, D., Nogueira, R., Sivan-Hoffmann, R., 2021. Evaluation of artificial intelligence–powered identification of large-vessel occlusions in a comprehensive stroke center. American Journal of Neuroradiology 42, 247–254.

Appendix A. Classification

Table A.8: Experiment of Resnet-18 pretrained with Videos, showing the performance of freezing a different amount of layers. Highlighted results show the best sensitivity and consistent results in the other metrics.

Frozen layers	Accuracy	Specificity	Sensitivity	AUROC
All	0.5000	0.7083	0.2727	0.4725
3	0.7609	0.7917	0.7273	0.7320
2	**0.7391**	**0.7917**	**0.6818**	**0.7405**
1	0.7391	0.7917	0.6818	0.8125
None	**0.6739**	**0.8750**	**0.4545**	**0.7850**

Appendix B. nnDetection configuration

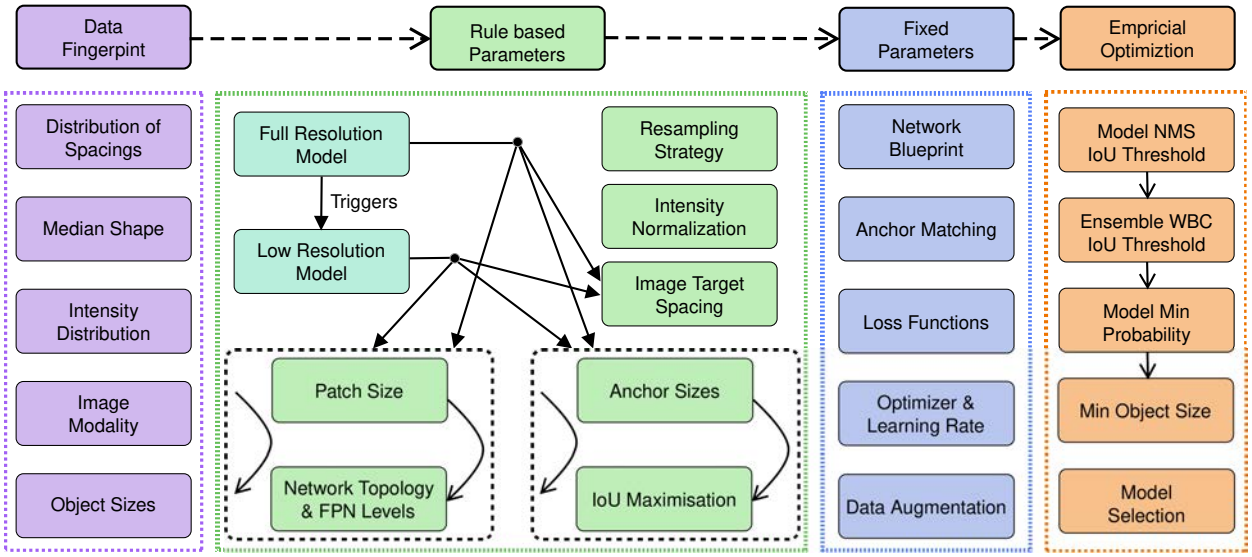


Figure B.12: Parameters configuration of nnDetection. Taken from Baumgartner et al. (2021).

Medical Imaging and Applications

Master Thesis, June 2023

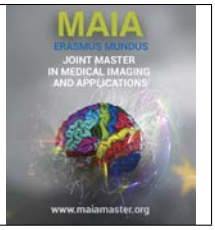


Image Transformers for Multi-view Lesion Detection in Mammography

Habtamu Tilahun Mekonnen, Robert Martí

Universitat de Girona, Girona, Spain

Abstract

In recent years, some approaches based on Deep Neural Networks (DNNs) have been developed with the aim of achieving optimal clinical outcomes through the use of multi-view mass detection techniques in Full Field Digital Mammograms (FFDMs). Convolutional Neural Networks (CNNs) have been the dominant method for implementing these techniques. However, transformer models which are mainly based on the Detection Transformer (DETR) have emerged as a viable alternative to CNNs. In this study, we investigate the benefits of using transformers for multi-view mass detection in mammographic images, incorporating information from the left and right mammograms as multi-channel input images. We propose two multi-view mass detection approaches that involve merging a mammogram with a lesion and a healthy contralateral image from the OPTIMAM Mammography Image Database (OMI-DB). One of the approaches incorporates a difference image, while the other approach does not. DETR and Deformable DETR with ResNet-50 as a feature extractor are used for the detection process. The best multi-view technique achieved a True Positive Rate (TPR) of 87.2% at 0.8 False Positives per Image (FPpI) and an area under the Free Receiving Operating Characteristic (FROC) curve of 79.9% on FFDMs from Hologic scanners. Furthermore, we present a single-view approach that demonstrates a significant improvement compared to the multi-view approaches. This approach achieves a TPR of 91.2% at 0.8 FPpI with an area under the FROC of 83.1%. The single-view models trained on Hologic images are then used for inference without fine-tuning. This inference is conducted on smaller datasets that contain FFDMs from the GE scanner, Siemens scanner, and another publicly available dataset known as VinDr-Mammo (IMS scanner). The best performance is achieved on the images from GE scanner with a TPR of 94.2% at 0.8 FPpI and an area under the FROC of 84.9%.

Keywords: Mammography, Breast Cancer, Computer aided detection, Deep learning, Transformers

1. Introduction

With an estimated 2.3 million new cases (11.7%) and 684,996 deaths worldwide in 2020, breast cancer was the most frequently diagnosed cancer and the leading cause of cancer deaths in women (Sung et al., 2021). The likelihood of breast cancer being successfully treated increases with early detection of lesions in the smallest size. Different imaging techniques can be used to diagnose breast cancer. Although its sensitivity considerably decreases with increasing breast density (Kolb et al., 2002), mammography is the standard imaging technique for breast cancer screening due to its fast acquisition and cost-effectiveness (Sree et al., 2011). This technique utilizes low-energy X-rays to build an image of the breast called a mammogram. For each breast, two different views are used to detect suspicious lesions like masses which appear as a well-defined or irregularly shaped area standing out from the surrounding tissue. The cranio-caudal (CC) view is taken from top to bottom direction, while the medio-lateral oblique (MLO) view is a side view from a specific angle. Although early cancer detection using screening mammography has decreased the number of women who have died of breast cancer by about 40%, manual examination of mammograms is a challenging task that depends on experience and fatigue level of the radiologist (Wang et al., 2014). Computer-Aided Detection (CADe) systems have been developed with the goal of providing radiologists with an additional perspective, assisting them in improving the accuracy of detection and localization of masses and other abnormal-

ities. These schemes aim to reduce significant variability among different radiologists. As deep learning algorithms have advanced significantly over the last decade, Convolutional Neural Networks (CNNs) substantially impacted the development of CADe systems. CNNs can focus on local features while retaining spatial relationships, making them an integral component of modern AI-based medical imaging systems. However, the inception of Vision Transformers (ViTs) in the work of Dosovitskiy et al. (2020) made researchers aware of CNNs primary drawback, their failure to understand global context or relationships that span across a large spatial area (Liu et al., 2022). Segmentation and classification being the most affected areas, transformers have significantly influenced all fields of medical imaging tasks (Shamshad et al., 2022). However, very little work has been done to enhance the performance of medical image detection by using transformer-based approaches, which mostly use the Detection Transformer (DETR) (Zhu et al., 2020).

Unlike radiologists, who combine information from multiple views of a mammogram to make diagnostic conclusions, most CADe schemes only use a single view, which has limited clinical utility (Jones et al., 2023). Research interest has recently increased in the development of multi-view CADe systems, which aggregate information extracted from different views, for lesion detection in mammography. These approaches perform either bilateral, ipsilateral or simultaneously both bilateral and ipsilateral analysis of mammographic images to emulate the radiologists' reading practice.

This study aims to investigate the benefits of using transformers for multi-view mass detection in mammographic images, incorporating information from the left and right mammograms as multi-channel input images, and to evaluate and compare their performance to single-image transformers and another traditional object detection method. Additionally, we analyzed the performance of transformers across different attention mechanisms and datasets obtained using different scanners.

The remaining of this paper is organized as follows: Section 2 summarizes the existing work on mass detection in mammography including the dataset and techniques used, and findings obtained. Section 3 explains the dataset, methods and detection networks used in this project. The results obtained from different experiments conducted in this study are presented in Section 4. Section 5 analyzes and interprets the findings in the context of existing literature and compares them with previous studies. Finally, Section 6 summarizes the main findings, and discusses potential future directions or further improvements.

2. State of the art

The literature on mammogram mass detection can be categorized into three distinct groups: pure image processing techniques, traditional feature-based approaches, and deep learning-based approaches. The research started with pure image processing approaches such as contrast enhancement filters (Petrick et al., 1996), relative image intensity-based techniques (Heath and Bowyer, 2000), region-growing techniques (Petrick et al., 1999), and template matching and gradient-orientation-analysis techniques (te Brake and Karssemeijer, 1999). Later, the study moved to machine learning-based approaches, which depend on handcrafted features to detect masses from mammogram images. Ke et al. (2010) presented a computer vision system for mass detection, relying on texture features. Their approach employed bilateral comparison to identify masses and determine the Region of Interest (ROI). Texture features such as fractal dimension and two-dimensional entropy were extracted from the ROI. The ROIs were then classified as either mass or normal using Support Vector Machines (SVM). The experiment was conducted on 106 mammograms, achieving a sensitivity of 85.11% at 1.44 false positives per image (FPpI). Rouhi et al. (2015) proposed a mass segmentation technique that uses region growing and cellular neural network methods. Genetic Algorithm (GA) was applied to select intensity histogram, shape, and texture features from the segmented images. These features were then used to classify masses into benign and malignant categories using various classifiers. The experiment conducted on the DDSM and MIAS datasets demonstrated a significant sensitivity of 96.87% for classification when employing the cellular neural-based segmentation technique. Mughal et al. (2017) applied mathematical morphology to extract and refine masses by generating a texture image using an entropy filter. Intensity, texture, and morphological features were extracted from the detected masses and classified using SVM, decision tree, K-Nearest Neighbor (KNN), and bagging tree classifiers. Among these classifiers, SVM achieved the best results. For the DDSM dataset, SVM achieved a sensitivity, specificity, and accuracy of 98.40%, 97.00%, and 96.9%, respectively. Similarly, for the MIAS dataset, the SVM classifier yielded a sensitivity, specificity, and accuracy of 98.00%, 97.00%, and 97.5%, respectively. Punitha et al. (2018) employed an improved variant of the region growing technique along with the dragon-fly optimization method to segment masses. Texture features extracted from the detected masses were classified using a Feed-Forward Network. By using 146 malignant cases and 154 benign cases from the DDSM, their approach achieved a sensitivity of 98.1% and a specificity of 97.8%. Lbachir et al. (2021) proposed a mass detection and diagnosis system comprising four sequential steps. In the first stage, the image undergoes preprocessing to enhance its contrast and eliminate any unwanted noise. Subsequently, the proposed Histogram Regions Analysis-based K-means (HRAK) algorithm is employed to segment abnormalities. In the

third step, texture and shape features, along with the bagged trees classifier, are used to decrease false positives. Finally, the abnormalities are classified as malignant or benign using the SVM. Their approach achieved a TPR of 93.15% at 0.467 FPpI on the MIAS dataset and a TPR of 90.85% at 0.65 FPpI on the CBIS-DDSM dataset.

In recent years, there have been notable advancements in the performance of CADe systems. These improvements primarily stem from the use of various promising deep-learning models such as Convolutional Neural Networks (CNNs), transfer learning techniques, and deep learning-based object detection models. Ribli et al. (2018) employed the DDSM database, comprising 2620 mammograms obtained from scanned films, to train a Faster R-CNN model. In their work, the performance of the model was evaluated on the INbreast dataset, achieving 90.0% TPR at 0.30 FPpI. Agarwal et al. (2020) assessed the performance of deep learning technique on the massive mammography dataset (OMI-DB) for the first time. Their study implemented a Faster R-CNN framework, which achieved a true positive rate (TPR) of 87.0% at a false positive per image (FPpI) value of 0.84. The evaluation was conducted on a subset of 7245 images obtained using Hologic scanners. Cao et al. (2020) introduced a new method for identifying breast masses in mammograms. They also presented a novel data augmentation technique to address the issue of overfitting caused by the limited dataset. Their augmentation technique employs local elastic deformation, which effectively improved the performance of their model. However, it should be noted that this technique takes longer to compute compared to traditional augmentation methods. To enhance the contrast between the breast mass and its surrounding area, they employed a combination of the truncation normalization method and adaptive histogram equalization. They utilized an improved version of the RetinaNet called Feature Selective Anchor-Free (FSAF) (Zhu et al., 2019) for mass detection, achieving a TPR of 93.0% TPR at 0.50 FPpI on the INbreast dataset. Su et al. (2022) introduced a double-shot model for simultaneous mass detection and segmentation, combining the YOLO and Local-Global (LOGO) architectures. This approach leveraged YoloV5 to accurately locate and crop the breast mass in mammograms and LOGO to segment the masses while ensuring the preservation of their original shape and position. The proposed model was assessed on the CBIS-DDSM and INbreast datasets, achieving a TPR of 95.7% and a mean average precision of 65.0% on the CBIS-DDSM dataset. Yu et al. (2023) introduced a patch-based approach for detecting breast masses, which consisted of three modules. Firstly, they employed an enhanced Deeplabv3+ model for pre-processing to remove the pectoral muscle. Secondly, they used a multiple-level thresholding segmentation method to extract candidate mass patches. Lastly, they employed trained deep learning models to classify these patches into either breast masses or background breast tissue. When evaluated on the CBIS-DDSM dataset, the method achieved a TPR of 87% at 2.86 FPpI. While on the INbreast dataset, the method achieved a TPR of 96% at a FPpI of 1.29.

The majority of approaches using Deep Neural Networks (DNNs) for mammogram analysis are primarily focused on single-view scenarios. Recently, some approaches based on DNNs have been developed with the aim of achieving optimal clinical outcomes through the use of multi-view techniques. Yan et al. (2021) introduced a multitasking framework for breast mass detection that combined CC and MLO mammograms. Their method employed YOLOv3 region proposals with a Siamese network that fuses patch-level mass vs. non-mass classification and dual-view mass matching. The performance of this approach was evaluated on the INbreast dataset, achieving a TPR of 96% at 0.26 FPpI. Yang et al. (2021a) introduced a tri-view mass detection method known as MommiNet. Their approach incorporated a Faster R-CNN network with a Siamese input module and a DeepLab network with a Siamese input module, enabling simultaneous ipsilateral and bilateral analysis on the DDSM dataset. Their approach achieved a recall rate of 0.8 at 0.5 FPpI, showcasing its effectiveness. Later, they presented MommiNet-v2 (Yang et al., 2021b) with a new high-resolution network (HRNet)-based architecture, with the goal of learning the symmetry and geometry constraints and fully aggregating the information from all views for accurate mass detection. This method achieved a recall rate of 0.83 at 0.5 FPpI on the DDSM dataset, outperforming the original MommiNet. Liu et al. (2021) introduced a multi-view mammogram mass detection framework called Anatomy-aware Graph Convolutional Network (AGN). The proposed framework comprised three primary modules: (i) the Bipartite Graph Convolutional Network (BGN), which captured the intrinsic geometric and semantic relations of ipsilateral views; (ii) the Inception Graph Convolutional Network (IGN), which modeled the structural similarities of bilateral views; and (iii) the Correspondence Reasoning Enhancement Module, aimed at enhancing the representation power of features. The performance of their approach was assessed on both the DDSM and private datasets, resulting in a recall rate of 87.6% at 0.5 FPpI for the DDSM dataset and 82% at 0.5 FPpI for the private dataset.

The use of transformer-based approaches in breast imaging and mass detection research is still in its early stages, as evidenced by their recent appearance in the literature. Chen et al. (2022) proposed a transformer-based multi-view approach having local and global transformer blocks of learning patch relationships independently and jointly within four mammograms obtained from two different views (CC/MLO) of the left and right breasts. The proposed technique was eval-

uated on a private dataset consisting of 949 sets of mammograms, including 470 malignant and 479 benign cases. The results showed that their approach outperformed state-of-the-art multi-view CNNs, with an AUC of 0.818. Betancourt Tarifa et al. (2023) proposed transformer-based models on the OMI-DB dataset for the first time, employing the Swin transformer as a backbone multiscale feature extractor. By combining detection predictions from the transformer and convolutional models, their method outperformed the state-of-the-art method, with a TPR of 78.1% at 0.1 FPpI on a subset of 7626 images obtained using Hologic scanners.

3. Material and methods

3.1. Dataset

3.1.1. OPTIMAM mammography database (OMI-DB)

OMI-DB (Halling-Brown et al., 2021) is a vast mammography image database containing over 2.5 million images from 173,319 women collected from three UK breast screening centers. The database comprises both unprocessed and processed FFDMs stored in DICOM format. The images are accompanied by expert-drawn ROIs that indicate the location and size of lesions, along with other relevant attributes. Additionally, the database includes clinical data related to the images, such as screening history, previous occurrences of cancer, biopsy results, and surgical procedures. The OMI-DB encompasses images captured by various scanner manufacturers, including Hologic Inc., Siemens, Philips, and General Electric (GE) Medical Systems. Each case in the database offers two standard views, namely the CC and MLO views, for both breasts.

Only the processed FFDMs with finding annotations are used in this work. The OMI-DB dataset contains various breast abnormalities, including masses, calcifications, architectural distortions, focal asymmetries, or their combinations. Since mass detection is the objective of this work, only cases with both detected masses and no abnormalities are taken into consideration. The training and validation of our single-view and multi-view mass detection approaches are conducted using images from Hologic Inc. scanners, as they constitute the majority of images in the dataset. A total of 3614 FFDMs with detected masses (positive images only) collected from 1912 patients were employed to train and validate our single-view technique. For testing purposes, we used 90 images with mass findings acquired using Siemens scanners from 50 patients, as well as 104 images with mass findings acquired using GE scanners from 55 patients. The multi-view technique in this study employed a total of 6576 images collected from 1644 patients, using Hologic scanners. Among these images, there are 3288 images with detected masses, each paired with their corresponding contralateral images (representing normal images from the opposite breast). The images were selected with a thorough examination to make sure that any images containing artifacts or undesirable elements (like implants) were excluded from the analysis.

3.1.2. VinDr-Mammo

VinDr-Mammo (Pham et al.) is a large-scale dataset of FFDM consisting of 5,000 four-view exams (equivalent to 20,000 DICOM images) collected from two Vietnamese hospitals, namely Hospital 108 (H108) and Hanoi Medical University Hospital (HMUH). The acquisition of the images was carried out using three distinct scanners: Siemens, IMS, and Planmed. The database provides an overall assessment of the breast through Breast Imaging Reporting and Data System (BI-RADS) categories (ranging from 1 to 5) and breast density levels (classified as A, B, C, or D). Furthermore, it offers extensive lesion-level annotations by marking bounding rectangles around breast abnormalities such as masses, calcifications, asymmetries, and architectural distortions. These annotations only target abnormalities classified as BI-RADS 3, 4, or 5, indicating the need for further examination and follow-up. In this study, 158 FFDMs acquired using IMS scanners from 78 patients are used for testing the single-view technique. The reason for using only 78 cases is that there are no additional mass cases available that were acquired using IMS scanners.

3.2. Data preparation and pre-processing

For the single-view approach, the dataset obtained from the Hologic scanners is divided into training and validation sets using an 80-20 ratio. The patient-wise division is performed, ensuring that all mammograms from a specific case are exclusively assigned to either the training or validation set. Specific information regarding the number of images is presented in Table 1. Furthermore, three additional datasets are used to test the single-view approach, as indicated in Table 2. These datasets include OMI-GE, representing OMI-DB images acquired from GE scanners; OMI-S, referring to OMI-DB images obtained from Siemens scanners; and VinDr-IMS, which denotes VinDr-Mammo images obtained from IMS scanners. Importantly, it should be noted that these three datasets are not used together for testing the method; rather, they are employed in separate experiments.

In the multi-view approach, 3288 single-channel images containing detected masses are merged with their respective normal images from the opposite breast, yielding a total of 3288 multi-channel images. It is worth noting that only 3288 images are utilized due to the lack of contralateral images (with the same episode) for the remaining mass images. These multi-channel images are then divided into training and validation sets based on patients, following the same 80-20 ratio as the single-view approach.

Approach	Split	Cases	Images
Single-view	Training	1550	2890
	Validation	362	724
	Total	1912	3614
Multi-view	Training	1315	2630
	Validation	329	658
	Total	1644	3288

Table 1: Summary of training and validation dataset and its corresponding methodology. All images are acquired using Hologic scanners.

Dataset	Cases	Images
OMI-G	55	104
OMI-S	50	90
VinDr-IMS	78	158

Table 2: Description of the additional datasets used for testing a single-view approach.

The initial mammograms were in DICOM format, but they were converted to Portable Network Graphics (PNG) format for further usage. The datasets comprised images with high pixel resolutions ranging from around 64 μm to 108 μm and sizes between 2,000 and 4,000 pixels. To ensure that only relevant information was provided to the network, the mammograms were cropped to focus on the breast area of each image. This cropping process involved applying binary thresholding to the original image, and then extracting the largest connected component that represents the breast mask. The bounding box surrounding this mask was subsequently identified and used to crop the image, as depicted in Figure 1. Due to limitations in computational resources, the cropped images were subsequently downsampled to a pixel resolution of 200 μm. The images are normalized, and their intensity is rescaled to 8 bits.

(a) (b) (c)

Figure 1: Cropping breast area: **a** Original image **b** Breast mask **c** Cropped breast area

3.3. Methodology

Our study is structured into two main approaches: a single-view approach and a multi-view approach. In the sections below, the following are described: (i) the approaches investigated in this work; (ii) the object detection methods; (iii) the training process; and (iv) the evaluation metrics.

3.3.1. Single-view and multi-view approaches

The single-view approach identifies and locates masses within the breast using information obtained from a single mammographic image. It achieves this by replicating image information from the single channel into three channels.

The multi-view method uses information from both the patient's left and right mammograms, considering them as multi-channel input image, to identify masses. This is accomplished by registering the left and right mammograms, with one of them having a mass (abnormal) and the other without a mass (normal). For each pair of mammograms, we performed an affine registration using Elastix, ensuring a global alignment between the images. The right mammogram was employed as the fixed image, while the left mammogram was used as the moving image. It is important to note that both the moving and fixed images were acquired during the same visit (episode) and have identical view positions (either both are CC or both are MLO). The main configuration for this registration involved a 4-level multiresolution scheme, with mutual information used as the cost function, adaptive stochastic gradient descent employed as the optimizer, and a first-order B-spline utilized as the interpolator. During the final step of registration, where the pixel correspondences between the two mammograms are identified, the moving image is transformed through resampling, which necessitates interpolation. To ensure more accurate alignment of the overall breast shape, we opted for a B-spline of order 3 as the resampling interpolator. The registration process produces two main outputs: the registered image, which is the transformed version of the moving image, and the six affine transformation parameters that describe the spatial adjustments applied during the registration. In cases where the left mammogram (a moving image) contains a mass, we performed a point transformation. This involved applying the affine transformation parameters learned from the registration process using Transformix to adjust the coordinates of the ground truth bounding boxes. The purpose was to ensure that these coordinates align with the registered image, accurately representing the same anatomical or spatial features as shown in Figure 3. This is useful to perform evaluations or any subsequent analysis on the same spatial coordinate system established by the registration.

Two distinct multi-view mass detection techniques are employed in this study: one involves the difference image between the right and registered left mammograms, while the other does not utilize a difference image. Figure 2 illustrates the multi-view mass detection framework, highlighting these techniques.

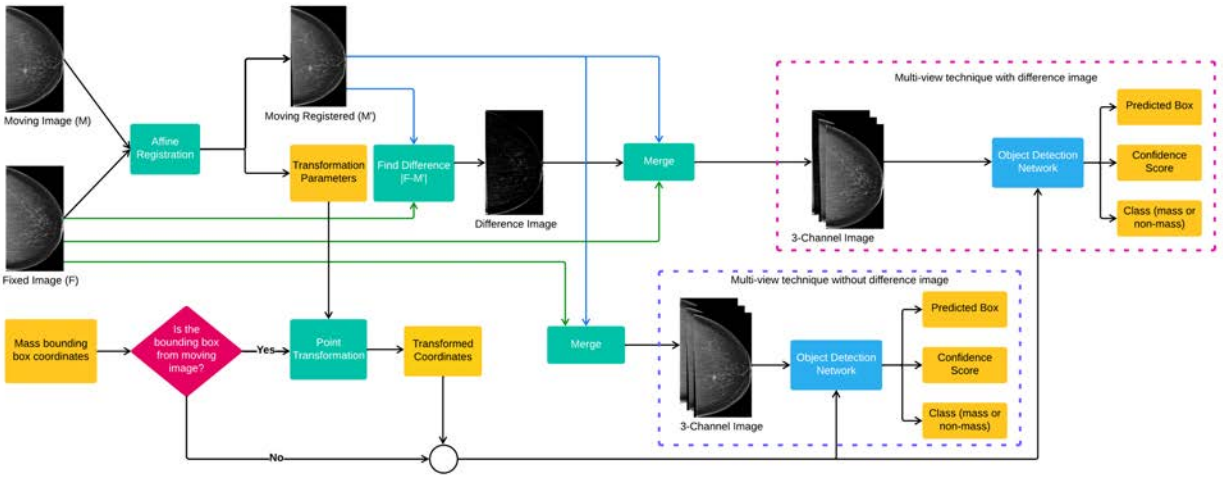


Figure 2: Multi-view mass detection framework

Multi-view technique with difference image

This technique involves combining the fixed image, the moving registered image, and their absolute difference to create a composite 3-channel image. This composite image will then serve as the input for the detection network. Hence, the process of detecting masses involves integrating data from both abnormal and healthy breasts, along with highlighted dissimilarities or changes between them.

Multi-view technique without difference image

This technique involves replicating the information from an image containing the mass twice and merging it with the information from a healthy contralateral image. The resulting image consists of three channels, where two channels contain the replicated information from the image with the mass, and the remaining channel contains information from the healthy image, with the aim of giving more importance to the image with the lesion, under the hypothesis that this could improve the detection results. This 3-channel image is then fed into the object detection network for the detection of masses in the abnormal breast.

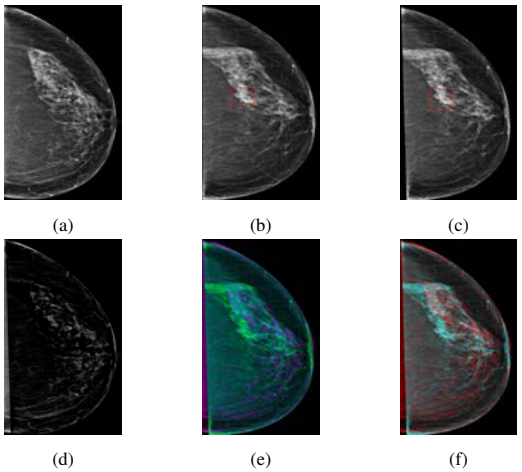


Figure 3: Image registration and data preparation for the multi-view mass detection approach: (**a**) Fixed image (F) representing the right mammogram (**b**) Moving image (M) representing the left mammogram with the ground truth bounding box, which indicates the detected mass (**c**) Moving registered (M') with the transformed bounding box (**d**) Difference image ($|F - M'|$) (**e**) 3-channel image created by merging F, M' and $|F - M'|$ (**f**) 3-channel image generated by duplicating M' in two channels and merging them with F

3.3.2. Object detection methods

In this study, we used three object detection networks, namely Faster R-CNN, Detection Transformer (DETR), and Deformable Detection Transformer (DDETR). The primary purpose of employing Faster R-CNN was to establish a baseline for our datasets. All three methods were used with a convolutional backbone.

Faster R-CNN

Girshick et al. (2014) proposed a Region-based Convolutional Neural Network (R-CNN) which gained a lot of interest in the computer vision community. The idea of R-CNN was to use a Selective Search (SS) approach to propose around 2000 ROIs, which were then fed into a CNN to extract features. These features were used to classify the images and determine their object boundaries using SVM and regression methods. R-CNN was quickly followed by Fast R-CNN (Girshick, 2015), a faster and better approach for object detection. Fast R-CNN used an ROI pooling approach, which shares the features across the whole image and uses a modified form of the spatial pyramid pooling method to extract features in a computationally efficient way. The problem with Fast R-CNN is that it is still slow because it needs to perform SS which is computationally time-consuming. This shortcoming led researchers to come up with Faster R-CNN (Ren et al., 2016), where the Region Proposal Network (RPN) is used as a region pro-

poser without the need for any external mechanism like SS. In Faster R-CNN, the input image is fed into the CNN, and the resulting feature map is given to the RPN, where 9 region boxes (anchors) of different scales and aspect ratios are used for generating region proposals, eliminating the need for image pyramids. The RPN generates a set of proposals, with each proposal assigned a probability score indicating its likelihood of being an object, along with the corresponding class or label of the object. After this, ROI pooling and then an upstream classifier and bounding box regressor are used, similar to Fast R-CNN. Figure 4 illustrates the RPN and the overall pipeline of the Faster R-CNN model.

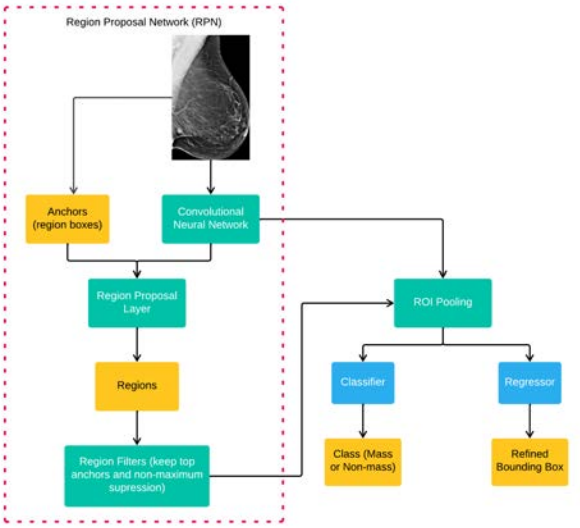


Figure 4: Flow chart of the Faster R-CNN, depicting both the region proposal network (RPN) and the overall pipeline.

DETR

Two-stage object detection architectures (like Faster R-CNN) employ hand-crafted components such as anchor generation and Non-maximum Suppression (NMS), which puts them behind the desired level in terms of speed. Carion et al. (2020) designed the DEtection TRansformer (DETR), a much faster technique trained end-to-end with a set loss function that performs bipartite matching between predicted and ground-truth objects, using a transformer encoder-decoder architecture. The DETR model consists of a CNN backbone (ResNet), which learns a 2D representation of an input image and produces a set of lower-dimensional features. These features are then flattened into a 1-dimensional structure and added to a positional encoding, which is fed into a transformer encoder. Learned positional embeddings, also called object queries, which represent N (where N is set to be significantly larger than the typical number of objects in an image) different learnable objects, are passed to the decoder part of the architecture to be learned with additional attention. Each output embedding of the decoder is then passed into a shared Feed

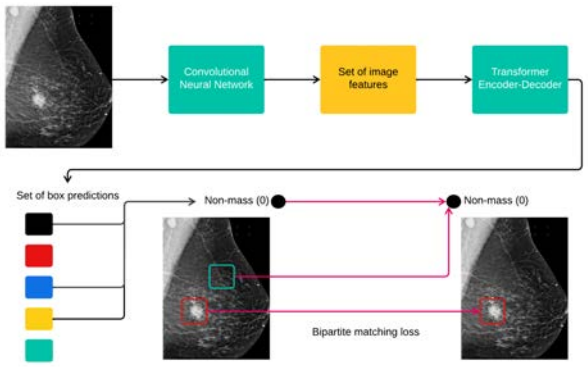


Figure 5: Schematic representation of DETR, showing the process of predicting the final set of detections by combining a CNN with a transformer architecture

Forward Network (FFN), a 3-layer perceptron with a ReLU activation function and hidden dimension d, that predicts either a detection (class and bounding box) or an $\emptyset$ (no object) class. The loss is calculated by computing the bipartite matching loss as shown in Figure 5.

Deformable DETR

DETR eliminates the necessity for hand-crafted elements like anchor generation, but it also faces challenges such as limited feature spatial resolution, extensive data requirements for training, and slow convergence during training (resulting in extended training periods). Due to the deficit of transformer components in processing image feature maps, DETR has a relatively poor performance in detecting small objects. Deformable DETR (Zhu et al., 2020) is proposed to address these problems by combining sparse spatial sampling of deformation convolution and the relation modeling capability of the transformer. Deformable DETR makes use of multi-scale feature maps to detect objects at different scales, especially small objects. It also introduces a deformable attention module, which only attends to a small set of key sampling points around reference points, regardless of the spatial dimension of feature maps. Hence, it brings down the complexity from quadratic (in the case of DETR) to linear, saving a lot of computation cost. However, it is worth mentioning that this computational cost saving is not always true, as it depends on the underlying implementation of the deformable attention module.

3.3.3. Network Training

The detection frameworks were trained using COCO pre-trained Resnet-50 as the convolutional backbone. Microsoft Common Objects in Context (COCO) is a widely used benchmark dataset for the tasks of object detection, instance segmentation, and image captioning research. It contains 328,000 images of everyday objects and humans with bounding box and segmentation

mask annotations of 91 object categories.

To train Faster R-CNN and DETR, we used Detectron2 (v0.6), a PyTorch-based open-source library introduced by Facebook AI Research (Wu et al., 2019). This library provides state-of-the-art object detection and segmentation algorithms. For training DDETR, we employed MMDetection (v2.28.2), another PyTorch-based open-source object detection toolset, presented by Chen et al. (2019). Both Detectron2 and MMDetection offer a wide range of COCO-pretrained models, training and evaluation tools, and utilities to facilitate object detection research and applications. Given that our dataset is not natively supported in Detectron2, we employed Dataset APIs (DatasetCatalog and MetadataCatalog) to enable Detectron2 access our dataset in the form of a list of standard dataset dictionaries. These dictionaries, which are the inputs for the Faster R-CNN and DETR models, contain essential information about the image such as file name, height, width, ID, and annotations. Furthermore, for training DDETR, we restructured our dataset into the COCO annotation format, which is commonly used in object detection tasks. The Faster R-CNN and DETR models were trained using a single NVIDIA A30 24 GB GPU, whereas the DDETR model was trained using a single NVIDIA A40 48 GB GPU.

In the Faster R-CNN algorithm, the anchors are generated at a specified pixel location. These anchors are defined by a base size of 128 pixels, three aspect ratios (0.5, 1.0, and 2.0), and five different scales (0.1, 0.2, 0.5, 1.0, and 2.0), resulting in a total of 15 anchors. Afterward, the detected bounding boxes undergo a non-maximum suppression (NMS) technique with a threshold of 0.05. This process ensures that only the most prominent and non-overlapping bounding box predictions remain, effectively representing masses found in the mammogram. The initialization of anchor boxes and NMS is based on the suggestion made by Agarwal et al. (2020).

Data Preprocessing

The images were adjusted in size, with the height and width not exceeding 800 and 1333 pixels, respectively, while maintaining the original aspect ratio as suggested in Betancourt Tarifa et al. (2023). However, due to the memory access pattern of deformable convolution, which can vary spatially and temporally and result in larger memory usage (Ahn et al., 2020), the maximum height and width for DDETR were limited to 600 and 1000 pixels, respectively. Furthermore, the pixel values were normalized to have a mean of zero and a standard deviation of one.

Data Augmentation

As suggested by Agarwal et al. (2020), during the training process of the Faster R-CNN-based detector, we employed only horizontal flip as a data augmentation. However, when training DETR and DDETR, we incorporated various data augmentation techniques to enhance the training process. Every image was subjected to one of the following techniques with a 50% probability: (i) horizontal flip, (ii) random crop, (iii) contrast transformation using magnitude values of [0.4, 0.8, 1.5], and (iv) brightness transformation using magnitude values of [0.3, 0.7, 1.3]. These techniques were suggested by Boutancort et al. (2023). During the training of Faster R-CNN and DETR models, the Detectron2 framework used its Dataset Mapper module to implement data augmentation. On the other hand, for training DDETR, the MMDetection framework's AutoAugment class was employed to incorporate data augmentation techniques.

Training Hyperparameters

All models underwent training in batches consisting of two images. The Faster R-CNN model was trained using the stochastic gradient descent (SGD) optimizer, while both DETR and DDETR models were trained using the ADAMW optimizer. Performance monitoring and early stopping were conducted using the mean Average Precision (mAP) calculated with an IoU threshold of 0.5. Training parameters used to finetune the detection models are described in Table 3.

3.3.4. Evaluation Metrics

To evaluate the performance of detection methods, free receiver operating characteristic (FROC) curves are used. These curves provide information about the True Positive Rate (TPR) of the detected masses in relation to the average number of False Positives per Image (FPpI). In order to plot the FROC, we varied the threshold of confidence probability (objectness score) generated by the network. Only the predicted boxes with confidence probabilities greater than or equal to a specified threshold are taken into account for each threshold value. A mass was considered as a true positive (TP) when the intersection over union (IoU) between the predicted box and the ground truth box was 10% or higher according to the criterion established in Agarwal et al. (2020). Masses that were not detected by the model are considered as false negatives (FNs), whereas all other predicted boxes with an IoU less than 10% are considered as false positives (FPs) for the image. When dealing with mammograms containing multiple masses, the IoU is computed individually for each ground truth box. The True Positive Rate (TPR) is calculated by using equation 1, where TP represents the total number of true positives, and FN represents the total number of false negatives in validation or testing images.

$$\text{TPR} = \frac{TP}{TP + FN} \tag{1}$$

The calculation of FPpI involves summing the count of false positives (FPs) in each image and dividing this sum by the total number of images. The FROC curves

Detection Method	Approach	Learning rate	Epochs
Faster R-CNN	Single-view	5×10^{-5}	32
	Multi-view with difference image	2.5×10^{-5}	35
	Multi-view without difference image	1.25×10^{-5}	35
DETR	Single-view	2.5×10^{-5}	32
	Multi-view with difference image	2.5×10^{-5}	35
	Multi-view without difference image	2.5×10^{-5}	30
DDETR	Single-view	1.25×10^{-5}	28
	Multi-view with difference image	1.25×10^{-5}	28
	Multi-view without difference image	1.25×10^{-5}	28

Table 3: Training Parameters

were analyzed to obtain two performance measures: (i) the True Positive Rate (TPR) at a False Positive per Image (FPpI) value of 0.8, and (ii) the Area Under the Curve (AUC) within the range of FPpI values from 0 to 1. Both of these performance measures are in line with the ones in Agarwal et al. (2020).

4. Results

This section presents the performance of the trained mass detection models in four separate parts. Initially, we showcase the results obtained using a single-view approach for different datasets. The subsequent two sections present the results achieved through a multi-view technique, one involving a difference image and the other without it. Finally, comparisons between single-view and multi-view approaches are presented in terms of AUC and TPR differences.

4.1. Mass detection with single-view technique

The FROC curves for various datasets employed in this technique are reported in Figure 6, Figure 7, and Figure 8, illustrating the effectiveness of Faster R-CNN, DETR, and DDETR models, respectively. Performance results of these models are also presented in Table 4. In a study conducted by Agarwal et al. (2020), they reported a TPR of 0.87 at 0.84 FPpI. In comparison, our baseline model achieved a TPR of 0.877 at 0.8 FPpI. These results demonstrate that we were successful in replicating the method and results of Agarwal et al. (2020). Therefore, it can serve as a baseline for comparison with our proposed methodologies.

It can be observed that DETR and DDETR outperformed our baseline model on all datasets except VinDr-IMS. Among the datasets used for evaluation, Faster R-CNN exhibits the best performance on the Hologic images, DETR performs the best on the OMI-GE dataset, and DDETR shows superior performance on the OMI-S dataset. On the other hand, when it comes to the VinDr-IMS dataset, all models showcased the lowest performance. It is worth noting that the DETR model trained on the Hologic dataset exhibited a higher TPR on the OMI-GE dataset when compared to the validation set of Hologic. Similarly, the DDETR model trained on the Hologic dataset demonstrated a higher TPR on both the OMI-GE and OMI-S datasets compared to the validation set of Hologic.

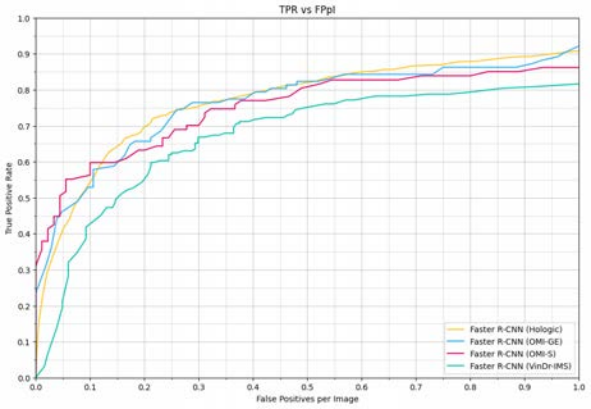


Figure 6: FROC curves of Faster R-CNN on different datasets with single-view approach

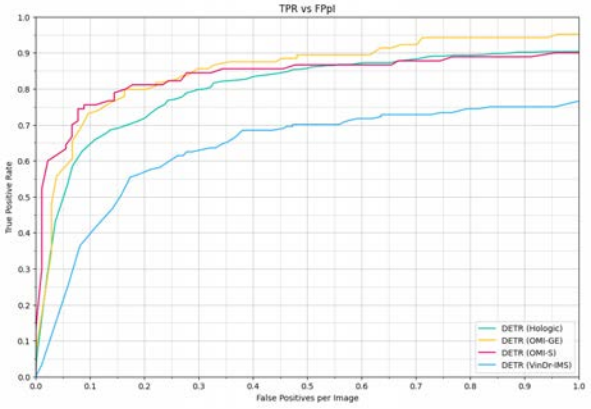


Figure 7: FROC curves of DETR on different datasets with single-view approach

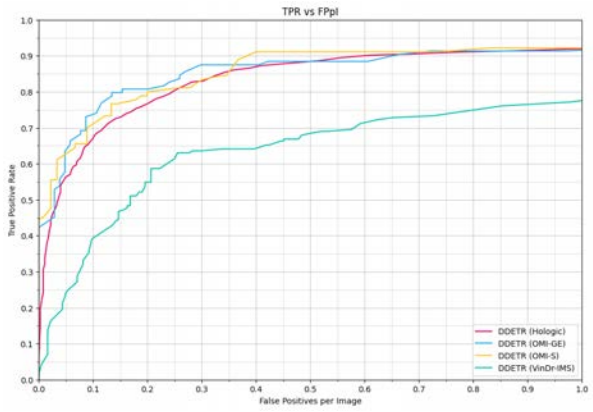


Figure 8: FROC curves of DDETR on different datasets with single-view approach

Model	Database	TPR at 0.8 FPpI	AUC
Faster R-CNN	Hologic	**87.7**%	**76.8**%
	OMI-GE	86.3%	76.1%
	OMI-S	83.9%	74.9%
	VinDr-IMS	79.4%	66.8%
DETR	Hologic	89.3%	80.0%
	OMI-GE	**94.2**%	**84.9**%
	OMI-S	88.9%	83.4%
	VinDr-IMS	74.7%	63.1%
DDETR	Hologic	91.2%	83.1%
	OMI-GE	91.3%	85.0%
	OMI-S	**92.2**%	**85.2**%
	VinDr-IMS	75%	62.9%

Table 4: Performance of trained models using a single-view approach on different datasets. The Hologic dataset serves as the validation set, whereas the OMI-S, OMI-GE, and VinDr-IMS datasets are employed as the test sets. Highest performance for each model is marked in bold.

In Figure 9, examples of single-view mass detection results given by the three detection models are shown. The predictions generated by the models are filtered, where only predicted boxes with a confidence probability of 0.3 or higher are kept. It is evident that DDETR robustly predicts masses without generating false positives, whereas both Faster R-CNN and DETR models exhibit instances of false positive detections.

4.2. Mass detection with multi-view technique involving a difference image

This approach is evaluated only on the Hologic dataset and the FROC curves of all experimented detectors are reported in Figure 10. When compared to the performance achieved through a single-view approach on the Hologic dataset, all detectors exhibit lower performance in this approach. Once more, the performance of DDETR followed by DETR outperforms that of the baseline model, as shown in Table5.

4.3. Mass detection with multi-view technique without involving a difference image

Figure 11 presents the FROC curves of all models trained using the multi-view technique without involving a difference image. Evaluation of the models is only conducted on the Hologic dataset, and in comparison to the performance obtained using a single-view approach, they achieved lower performance. However, there is a significant improvement in performance when compared to the multi-view technique that incorporates the difference image. As demonstrated in Table 5, the performance of DDETR followed by DETR surpasses that of the baseline model, similar to both the single-view technique and the multi-view technique involving the difference image.

4.4. Comparisons between single-view and multi-view approaches

For every combination of single-view and multi-view methods, the differences in AUC (ΔAUC) and TPR (ΔTPR) were assessed as shown in Table 6. For each detection model, the difference between the performance achieved by the multi-view technique involving a difference image and the single-view technique is significantly higher, indicating that the multi-view technique involving a difference image has the lowest performance. On the other hand, the difference between the performance achieved by the multi-view technique without involving a difference image and the single-view technique is comparatively lower, suggesting that the multi-view technique without involving a difference image has a relatively lower performance.

5. Discussion

In this study, multi-view mass detection in mammographic images, incorporating information from the left and right mammograms as multi-channel input images is investigated by using transformer-based detection heads. It has been demonstrated that incorporating the difference image as one of the input image channels leads to a reduction in detection performance. This decrease in performance could be attributed to several factors. The difference image contains pixel-level intensity variations between the mammogram with a mass and the healthy mammogram. By using the difference image as a separate channel, we may lose valuable information present in the original left and right mammograms. This loss of information can lead to a decrease in performance. The difference image may also introduce additional noise or irrelevant features into the input data. This can negatively impact the model's ability to learn meaningful patterns and discriminate between mass and

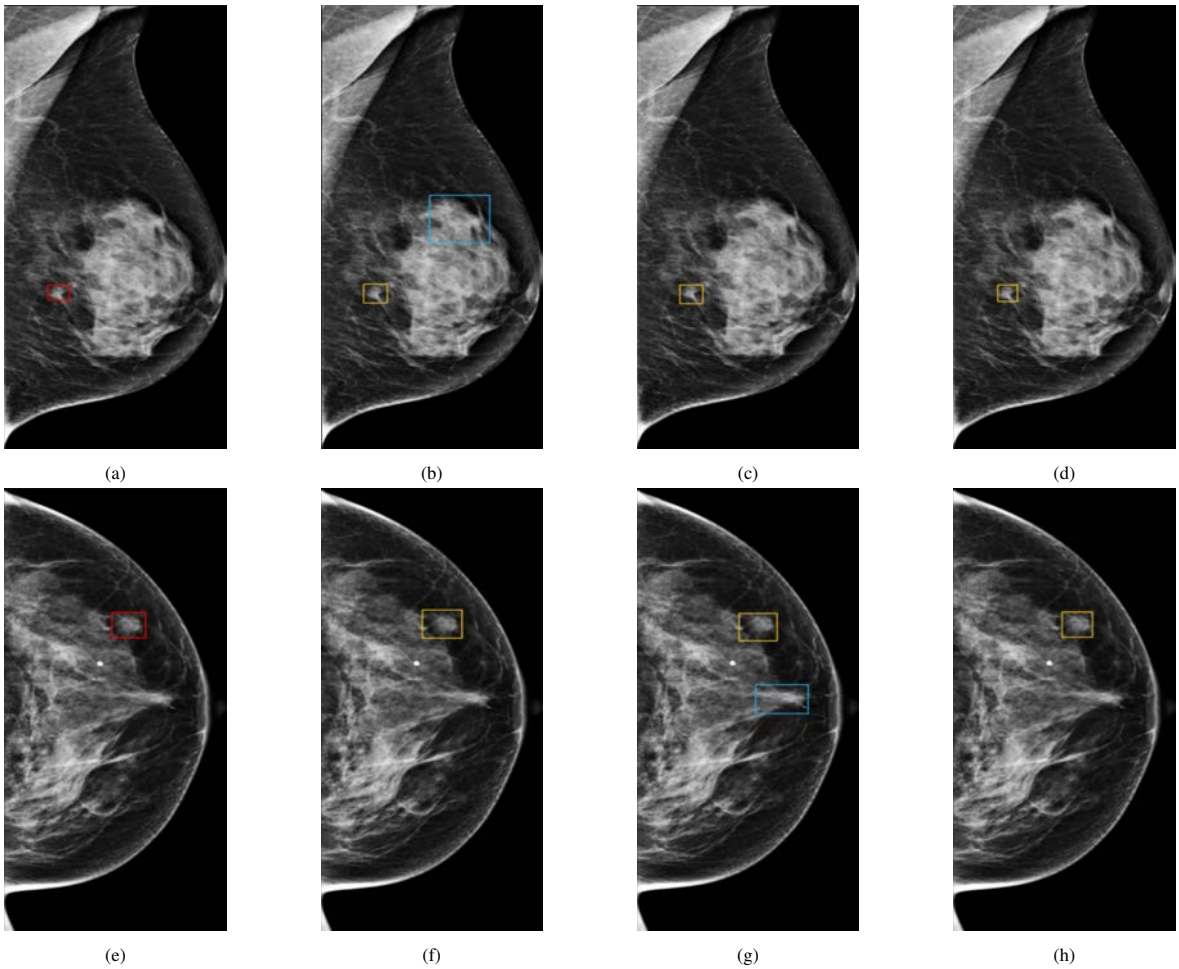


Figure 9: Examples of mass detection results obtained from Hologic images: **a** and **e** are images with groundtruth bounding box; **b** and **f** are images with predicted boxes using Faster R-CNN; **c** and **g** are images with predicted boxes using DETR; **d** and **f** are images with predicted box using DDETR. TP detection is represented by yellow bounding box, FP detection is represented by blue bounding box.

Model	Approach	TPR at 0.8 FPpI	AUC
Faster R-CNN	Multi-view with difference image	72.5%	63.5%
	Multi-view without difference image	83.0%	72.7%
DETR	Multi-view with difference image	74.7%	64.9%
	Multi-view without difference image	85.0%	76.1%
DDETR	Multi-view with difference image	77.7%	67.9%
	Multi-view without difference image	87.2%	79.9%

Table 5: Performance of trained models using a multi-view approach

healthy tissue regions. Furthermore, if the difference between the left and right mammograms is not consistent across different cases, the model may struggle to effectively utilize the difference image. This can result in a less informative representation for distinguishing between mass and healthy regions.

With the aim to give greater importance to the image with lesion, under the hypothesis that this could enhance detection results, an assessment is conducted on a multi-view technique that combines information from the image with a mass replicated in two channels, along with information from a healthy image represented in one channel. This technique potentially improved detection performance and demonstrated the significance of the lesion image. However, the performance of this technique is still below that of the single-view tech-

Model	Approach	ΔTPR	ΔAUC
Faster R-CNN	Multi-view with difference image	-15.2%	-13.3%
	Multi-view without difference image	-4.7%	-4.1%
DETR	Multi-view with difference image	-14.6%	-15.1%
	Multi-view without difference image	-4.3%	-3.9%
DDETR	Multi-view with difference image	-13.5%	-15.2%
	Multi-view without difference image	-4.0%	-3.2%

Table 6: Difference in TPR and AUC of multi-view approaches compared to the single-view approach

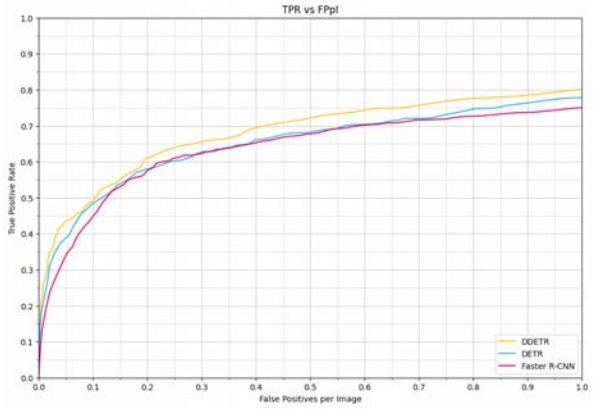


Figure 10: FROC curves of detectors using the multi-view technique with a difference image

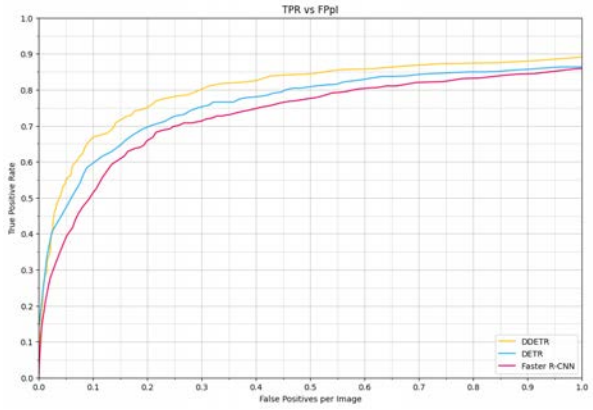


Figure 11: FROC curves of detectors using the multi-view technique without a difference image

nique. The healthy breast channel might contain features or patterns that are unrelated to the task of mass detection. As a result, the performance of the model might decrease because it is attempting to consider information that is not beneficial for identifying masses. Implementing a multi-view approach based on ensemble methods might improve the performance of mass detection models.

In this study, it has also been shown that the Faster R-CNN, DETR, DDETR models, pre-trained on COCO dataset, can be fine-tuned to detect masses in FFDMS. This demonstrated that both CNN and transformer-based models will benefit from pretrained weights (Caron et al., 2021; Raghu et al., 2019; Taher et al., 2021). In both single-view and multi-view approaches, DDETR outperformed both DETR and Faster R-CNN counterparts. This is due to the use of multi-scale feature maps by DDETR to detect objects at different scales, especially small objects like subtle or occult masses as OMI-DB contains masses with different conspicuity levels (Betancourt Tarifa et al., 2023). On the other hand, DETR and Faster R-CNN exhibit relatively poor performance in detecting small objects. Furthermore, DETR achieved lower performance compared to DDETR due to its high dependence on abundant training data, similar to Vision Transformer (ViT).

In the single-view technique, transformer-based detection heads performed better than Faster R-CNN during inference on OMI-GE and OMI-S without finetuning as demonstrated in Matsoukas (2021). In all detection models, lowest detection performance was obtained on VinDr-IMS dataset. This is because VinDr-IMS has low contrast compared to the Hologic dataset. Enhancing the contrast of VinDr-IMS using techniques like histogram matching might benefit the inference process.

In this study, transformer-based backbones such as Swin and deeper Convolutional backbones like ResNet-101 were not used. Exploring these as an alternative feature extractors could potentially enhance the detection performance. Additionally, incorporating novel transformer-based detection heads like Hybrid Deformable DETR (H-DETR) (Jia et al., 2023) may further improve the detection results.

6. Conclusions

This study explores the detection of masses in mammographic images using a multi-view approach, where information from both left and right mammograms is merged into a multi-channel input image. Two transformer-based detectors are used on the large-scale mammography dataset OMI-DB, and two different multi-view mass detection approaches are assessed,

one involving a difference image and the other excluding a difference image. It is indicated that incorporating the difference image as an input image channel results in a decline in detection performance. Furthermore, the study reveals that the conventional single-view approach exhibits superior performance compared to both of the proposed multi-view approaches.

The study demonstrated that transformer-based models, which were originally pre-trained on natural images, can be fine-tuned and effectively adapted for the purpose of detecting masses in mammograms. The transformer-based detectors showcased superior performance compared to their convolutional counterparts, and DDETR appeared to be the most robust model for detecting masses. Using the single-view model trained on Hologic images, inference is conducted on smaller datasets without the need for fine-tuning. This inference showcased improved performance on datasets with a contrast level similar to that of Hologic images.

Our future work will involve the implementation of a multi-view mass detection approach based on ensemble methods to effectively leverage the complementary information provided by different views, aiming to achieve improved accuracy and robustness in the detection process. It would be also interesting to explore transformer-based backbones as a multi-scale feature extractor and deeper convolutional backbones as they exhibited better performance than shallower backbones in Betancourt Tarifa et al. (2023). Furthermore, an investigation would be performed into novel transformer-based object detectors.

Acknowledgments

I am immensely grateful to Prof. Robert Marti, my supervisor, for his unwavering commitment to sharing his expertise, providing consistent guidance, support, and encouragement throughout this Master's Thesis. I would also like to express my appreciation to the MAIA program for granting me this invaluable opportunity.

I would like to extend my special thanks to the colleagues within the ViCOROB research group for their invaluable assistance in clarifying doubts, offering valuable suggestions, and exchanging insightful ideas.

References

Agarwal, R., Díaz, O., Yap, M.H., Lladó, X., Martí, R., 2020. Deep learning for mass detection in Full Field Digital Mammograms. Computers in Biology and Medicine 121, 103774. URL: https://www.sciencedirect.com/science/article/pii/S001048252030144X, doi:10.1016/j.compbiomed.2020.103774.

Ahn, S., Chang, J.W., Kang, S.J., 2020. An Efficient Accelerator Design Methodology for Deformable Convolutional Networks. URL: http://arxiv.org/abs/2006.05238, doi:10.48550/arXiv.2006.05238. arXiv:2006.05238 [cs].

Betancourt Tarifa, A.S., Marrocco, C., Molinara, M., Tortorella, F., Bria, A., 2023. Transformer-based mass detection in digital mammograms. Journal of Ambient Intelligence and Humanized Computing 14, 2723–2737. URL: https://doi.org/10.1007/s12652-023-04517-9, doi:10.1007/s12652-023-04517-9.

te Brake, G., Karssemeijer, N., 1999. Single and multiscale detection of masses in digital mammograms. IEEE Transactions on Medical Imaging 18, 628–639. doi:10.1109/42.790462. conference Name: IEEE Transactions on Medical Imaging.

Cao, H., 2020. Breast mass detection in digital mammography based on anchor-free architecture. URL: http://arxiv.org/abs/2009.00857, doi:10.48550/arXiv.2009.00857. arXiv:2009.00857 [cs, eess].

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-End Object Detection with Transformers. URL: http://arxiv.org/abs/2005.12872, doi:10.48550/arXiv.2005.12872. arXiv:2005.12872 [cs].

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging Properties in Self-Supervised Vision Transformers. URL: http://arxiv.org/abs/2104.14294, doi:10.48550/arXiv.2104.14294. arXiv:2104.14294 [cs].

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D., 2019. MMDetection: Open MMLab Detection Toolbox and Benchmark. URL: http://arxiv.org/abs/1906.07155, doi:10.48550/arXiv.1906.07155. arXiv:1906.07155 [cs, eess].

Chen, X., Zhang, K., Abdoli, N., Gilley, P.W., Wang, X., Liu, H., Zheng, B., Qiu, Y., 2022. Transformers Improve Breast Cancer Diagnosis from Unregistered Multi-View Mammograms. Diagnostics 12, 1549. URL: https://www.mdpi.com/2075-4418/12/7/1549, doi:10.3390/diagnostics12071549. number: 7 Publisher: Multidisciplinary Digital Publishing Institute.

Girshick, R., 2015. Fast R-CNN, in: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1440–1448. doi:10.1109/ICCV.2015.169. iSSN: 2380-7504.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. URL: http://arxiv.org/abs/1311.2524, doi:10.48550/arXiv.1311.2524. arXiv:1311.2524 [cs].

Halling-Brown, M.D., Warren, L.M., Ward, D., Lewis, E., Mackenzie, A., Wallis, M.G., Wilkinson, L.S., Given-Wilson, R.M., McAvinchey, R., Young, K.C., 2021. OPTIMAM Mammography Image Database: A Large-Scale Resource of Mammography Images and Clinical Data. Radiology: Artificial Intelligence 3, e200103. URL: http://pubs.rsna.org/doi/10.1148/ryai.2020200103, doi:10.1148/ryai.2020200103.

Heath, M., Bowyer, K., 2000. Mass detection by relative image intensity .

Jia, D., Yuan, Y., He, H., Wu, X., Yu, H., Lin, W., Sun, L., Zhang, C., Hu, H., 2023. DETRs with Hybrid Matching. URL: http://arxiv.org/abs/2207.13080, doi:10.48550/arXiv.2207.13080. arXiv:2207.13080 [cs].

Jones, M.A., Sadeghipour, N., Chen, X., Islam, W., Zheng, B., 2023. A multi-stage fusion framework to classify breast lesions using deep learning and radiomics features computed from four-view mammograms. Medical Physics doi:10.1002/mp.16419.

Ke, L., Mu, N., Kang, Y., 2010. Mass computer-aided diagnosis method in mammogram based on texture features, in: 2010 3rd International Conference on Biomedical Engineering and Informatics, pp. 354–357. doi:10.1109/BMEI.2010.5639515. iSSN: 1948-2922.

Kolb, T.M., Lichy, J., Newhouse, J.H., 2002. Comparison of the Performance of Screening Mammography, Physical Examination, and Breast US and Evaluation of Factors that Influence Them: An Analysis of 27,825 Patient Evaluations1. Radiology URL: https://pubs.rsna.org/doi/10.1148/radiol.2251011667,

doi:10.1148/radiol.2251011667. publisher: Radiological Society of North America.

Lbachir, I.A., Daoudi, I., Tallal, S., 2021. Automatic computer-aided diagnosis system for mass detection and classification in mammography. Multimedia Tools and Applications 80, 9493–9525. URL: https://doi.org/10.1007/s11042-020-09991-3, doi:10.1007/s11042-020-09991-3.

Liu, Y., Zhang, F., Chen, C., Wang, S., Wang, Y., Yu, Y., 2021. Act Like a Radiologist: Towards Reliable Multi-view Correspondence Reasoning for Mammogram Mass Detection. URL: http://arxiv.org/abs/2105.10160. arXiv:2105.10160 [cs].

Liu, Y., Zhang, Y., Wang, Y., Hou, F., Yuan, J., Tian, J., Zhang, Y., Shi, Z., Fan, J., He, Z., 2022. A Survey of Visual Transformers. URL: http://arxiv.org/abs/2111.06091. arXiv:2111.06091 [cs].

Matsoukas, C., Haslum, J.F., Söderberg, M., Smith, K., 2021. Is it Time to Replace CNNs with Transformers for Medical Images? URL: http://arxiv.org/abs/2108.09038, doi:10.48550/arXiv.2108.09038. arXiv:2108.09038 [cs].

Mughal, B., Sharif, M., Muhammad, N., 2017. Bi-model processing for early detection of breast tumor in CAD system. The European Physical Journal Plus 132, 266. URL: https://doi.org/10.1140/epjp/i2017-11523-8, doi:10.1140/epjp/i2017-11523-8.

Petrick, N., Chan, H.P., Sahiner, B., Helvie, M.A., 1999. Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms. Medical Physics 26, 1642–1654. doi:10.1118/1.598658.

Petrick, N., Chan, H.P., Sahiner, B., Wei, D., 1996. An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection. IEEE transactions on medical imaging 15, 59–67. doi:10.1109/42.481441.

Pham, H.H., Nguyen Trung, H., Nguyen, H.Q., . VinDr-Mammo: A large-scale benchmark dataset for computer-aided detection and diagnosis in full-field digital mammography. URL: https://physionet.org/content/vindr-mammo/1.0.0/, doi:10.13026/BR2V-7517.

Punitha, S., Amuthan, A., Joseph, K.S., 2018. Benign and malignant breast cancer segmentation using optimized region growing technique. Future Computing and Informatics Journal 3, 348–358. URL: https://www.sciencedirect.com/science/article/pii/S2314728818300679, doi:10.1016/j.fcij.2018.10.005.

Raghu, M., Zhang, C., Kleinberg, J., Bengio, S., 2019. Transfusion: Understanding Transfer Learning for Medical Imaging. URL: http://arxiv.org/abs/1902.07208. arXiv:1902.07208 [cs, stat].

Ren, S., He, K., Girshick, R., Sun, J., 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. URL: http://arxiv.org/abs/1506.01497, doi:10.48550/arXiv.1506.01497. arXiv:1506.01497 [cs].

Ribli, D., Horváth, A., Unger, Z., Pollner, P., Csabai, I., 2018. Detecting and classifying lesions in mammograms with Deep Learning. Scientific Reports 8. doi:10.1038/s41598-018-22437-z.

Rouhi, R., Jafari, M., Kasaei, S., Keshavarzian, P., 2015. Benign and malignant breast tumors classification based on region growing and CNN segmentation. Expert Systems with Applications 42, 990–1002. URL: https://www.sciencedirect.com/science/article/pii/S0957417414005594, doi:10.1016/j.eswa.2014.09.020.

Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S., Fu, H., 2022. Transformers in Medical Imaging: A Survey. URL: http://arxiv.org/abs/2201.09873, doi:10.48550/arXiv.2201.09873. arXiv:2201.09873 [cs, eess].

Sree, S.V., Ng, E.Y.K., Acharya, R.U., Faust, O., 2011. Breast imaging: A survey. World Journal of Clinical Oncology 2, 171–178. doi:10.5306/wjco.v2.i4.171.

Su, Y., Liu, Q., Xie, W., Hu, P., 2022. YOLO-LOGO: A transformer-based YOLO segmentation model for breast mass detection and segmentation in digital mammograms. Computer Methods and Programs in Biomedicine 221, 106903. URL: https://www.sciencedirect.com/science/article/pii/S0169260722002851, doi:10.1016/j.cmpb.2022.106903.

Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F., 2021. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: a cancer journal for clinicians 71, 209–249. doi:10.3322/caac.21660.

Taher, M.R.H., Haghighi, F., Feng, R., Gotway, M.B., Liang, J., 2021. A Systematic Benchmarking Analysis of Transfer Learning for Medical Image Analysis. URL: http://arxiv.org/abs/2108.05930, doi:10.48550/arXiv.2108.05930. arXiv:2108.05930 [cs, eess].

Wang, Z., Yu, G., Kang, Y., Zhao, Y., Qu, Q., 2014. Breast tumor detection in digital mammography based on extreme learning machine. Neurocomputing 128, 175–184. URL: https://www.sciencedirect.com/science/article/pii/S0925231213010163, doi:10.1016/j.neucom.2013.05.053.

Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R., 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Yan, Y., Conze, P.H., Lamard, M., Quellec, G., Cochener, B., Coatrieux, G., 2021. Towards improved breast mass detection using dual-view mammogram matching. Medical Image Analysis 71, 102083. doi:10.1016/j.media.2021.102083.

Yang, Z., Cao, Z., Zhang, Y., Han, M., Xiao, J., Huang, L., Wu, S., Ma, J., Chang, P., 2021a. MommiNet: Mammographic Multi-View Mass Identification Networks.

Yang, Z., Cao, Z., Zhang, Y., Tang, Y., Lin, X., Ouyang, R., Wu, M., Han, M., Xiao, J., Huang, L., Wu, S., Chang, P., Ma, J., 2021b. MommiNet-v2: Mammographic multi-view mass identification networks. Medical Image Analysis 73, 102204. URL: https://www.sciencedirect.com/science/article/pii/S1361841521002498, doi:10.1016/j.media.2021.102204.

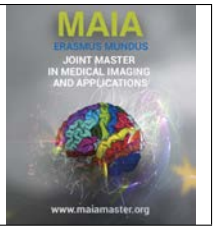
Yu, X., Wang, S.H., Zhang, Y.D., 2023. Multiple-level thresholding for breast mass detection. Journal of King Saud University - Computer and Information Sciences 35, 115–130. URL: https://www.sciencedirect.com/science/article/pii/S1319157822004049, doi:10.1016/j.jksuci.2022.11.006.

Zhu, C., He, Y., Savvides, M., 2019. Feature Selective Anchor-Free Module for Single-Shot Object Detection. URL: http://arxiv.org/abs/1903.00621, doi:10.48550/arXiv.1903.00621. arXiv:1903.00621 [cs].

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J., 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection. URL: http://arxiv.org/abs/2010.04159, doi:10.48550/arXiv.2010.04159. arXiv:2010.04159 [cs] version: 1.

Medical Imaging and Applications

Master Thesis, June 2023



MAM-E: Mammographic synthetic image generation with diffusion models

Ricardo Montoya del Ángel, Robert Martí Marly

ViCOROB Lab, Universitat de Girona, Girona, Spain

Abstract

Generative models have been used as an alternative data augmentation technique to counter the data scarcity problem faced in the medical imaging field. Diffusion models have gathered special attention due to their innovative generation approach, the high quality of the generated images and their relatively less complex training process compared with GANs. Still, the implementation of such models in the medical domain remains at early stages. In this work, we propose exploring the use of diffusion models for the generation of high quality full-field digital mammograms using state-of-the-art conditional diffusion pipelines. Additionally, we propose using stable diffusion models for the inpainting of synthetic lesions on healthy mammograms. We introduce *MAM-E*, a pipeline of generative models for high quality mammography synthesis controlled by a text prompt and capable of generating synthetic lesions on specific sections of the breast. Finally, we provide quantitative and qualitative assessment of the generated images and easy-to-use graphical user interfaces for mammography synthesis.

Keywords: generative models, mammography, stable diffusion

1. Introduction

Artificial intelligence has gained important attention in the last decade in essentially all aspects of human life. Thanks to the increasing data availability neural networks have played a key role on unveiling unsolved challenges, redefining AI research, and discovering new technological boundaries and applications.

A field that has attracted special recent attention is the generation of synthetic data, with the notable popularity of AI tools such as ChatGPT and DALL-E. Specifically in the imaging domain, generative models (GMs) started to gain notability in 2014 due to the impressive generative power of Generative Adversarial Networks (GANs). According to Yann LeCun, an important voice in the DL community, GANs were "..the most interesting idea in the last 10 years", as mentioned in his keynote at the Neural Information Processing Systems conference (NIPS) 2016 in Barcelona.

In the following years the appearance of new architectures and DL techniques, such as the rise of attention and transformers (Vaswani et al., 2017), further improved the generation capabilities and photorealism of the generated images in the natural imaging domain[1]. At the same time, researchers started to introduce these synthetic generation techniques into the medical imaging domain.

Contrary to natural images, medical images suffer from a data scarcity problem. Medical images are inherently more expensive than natural images due to their acquisition, processing and labeling procedure. Moreover, they are subject to more privacy and data protection concerns and, for some rare medical cases, images are difficult to find or suffer from underrepresentation, which leads to a subsequent data unbalance problem. All these issues dramatically reduce the volume of medical data available for the training of DL models, which limits the models performance and holds back the development of CAD systems, compared with non-medical imaging applications.

To counter this issue, GMs have been used to complement traditional data augmentation techniques and expand medical datasets, aiming to improve CAD mod-

[1] We refer as natural images to non-medical images, such as those included in large-scale datasets like ImageNet and LAION-B5. Other authors like Pinaya et al. (2022) have used this term.

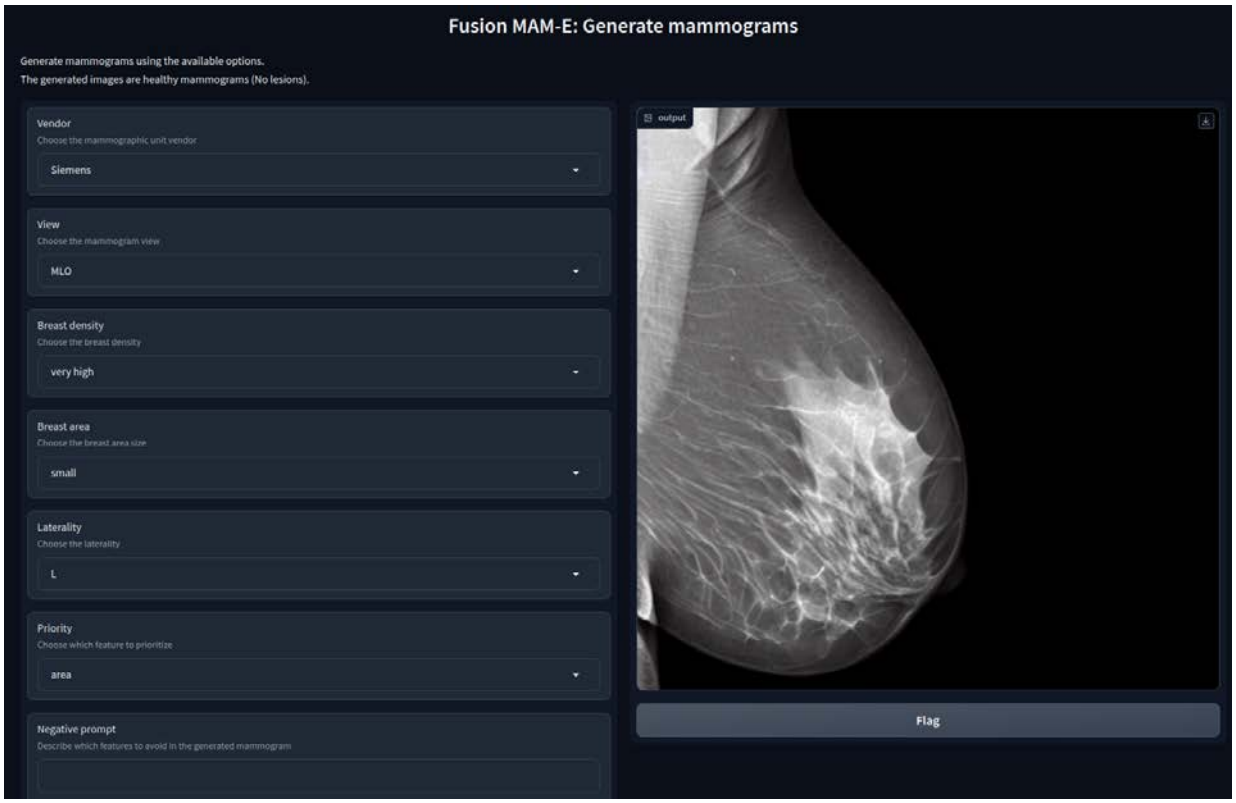


Figure 1: MAM-E: a synthetic mammogram generation tool.

els performance. Until a couple of years, GANs were the state-of-the-art (SOTA) for synthetic image generation tasks due to their high image quality and impressive photorealism. Nevertheless, some important limitations and drawbacks are inherent to these models. Due to its generator-discriminator architecture, GANs training is notoriously unstable and can be difficult to converge, as well as suffering from low diversity generation due to mode collapse issues (Kazerouni et al., 2023).

These issues make the use of GAN-like architectures challenging in some research domains. This is specially crucial for medical data, as medical diagnosis can highly depend on subtle changes in the organs appearance reflected in the images, changing the prediction of a CAD system (Müller-Franzes et al., 2022).

In 2021 diffusion models (DMs) captured the spotlight of the GMs community after the publication of OpenAI's belligerent article *Diffusion models beat GANs on Image Synthesis* by Dhariwal and Nichol (2021). Inspired by non-equilibrium thermodynamics, diffusion models rely on the idea that data distributions can be learned by iteratively destroying input information, adding certain noise, and then tasking a DL model to learn how to remove it in a denoising process, following a Markov chain.

Since the breakthrough of DMs, a great number of applications and research papers for natural images have been published to explore this new image generation principle. Results have shown promising improvements to the image generation task that continues to outperform GANs-like pipelines. Two main enhancements on traditional DM architectures are latent diffusion (LD), introducing the use of a latent space for higher image resolution, and stable diffusion models (SD) for additional input during training and inference for a more controlled generation process.

The medical image community has started to implement these improvements to generate high quality, high fidelity and realistic medical images, crucial characteristics for CAD systems development. Nevertheless, to the moment of publication of this work, the use of diffusion models in the medical imaging field continues at early stages. Even though a number of works have been published for the generation of several medical imaging modalities, such as brain MRI and chest X-ray, there is still no implementation of DM techniques for mammographic image synthesis.

1.1. Project description

The objective of this master thesis project is to explore the use of diffusion models for the generation of high-resolution mammographic images and to develop a synthesis pipeline using SOTA conditional diffusion models. This pipeline was developed using stable diffusion, a diffusion model technique that uses both conditioning, to control the image generation, and a latent

space to allow high-resolution without requiring large computational resources. The generated images are *for presentation*, meaning that their appearance and pixel intensities are meant for radiologist inspection, with the limitations on resolution and pixel depth inherent to the current state of diffusion pipelines.

The pipeline can be separated in two main tasks: healthy mammogram generation and lesion inpainting. For the first task, the generation process is controlled (or guided) using text conditioning with the description of the image using common mammography characteristics such as view position, laterality, breast density and breast area. For the second task we use an stable diffusion inpainting model designed to generate synthetic lesions in desired regions of the a mammogram.

We introduce *MAM-E*, a pipeline of generative models for high quality mammographic image synthesize, capable of generating images based on a text prompt, and also capable of generating lesions on a specific section of the breast. We selected the name after DALL-E, OpenAI's famous image generation tool for natural images presented by Ramesh et al. (2021), as we aimed to create a graphical user interface (GUI) similar to DALL-E to allow user personalization of the generated image based on customizable settings.

2. State of the art

2.1. Diffusion on medical imaging

Several relevant works have explored the implementation of diffusion models for synthetic medical images generation. Dorjsembe et al. (2022) proposed using the original pipeline of diffusion models on computer vision, introduced by Ho et al. (2020) called denoising diffusion probabilistic models (DDPM), for the generation of high-quality MRI of brain tumors, being the first attempt to investigate diffusion models for 3D medical images. This vanilla model was able to reproduce SOTA results, outperforming the baseline models based on 3D GANs.

A further improvement for synthetic brain MRI generation was presented by Pinaya et al. (2022), who used a Latent Diffusion model (LDM) to generate high-resolution 3D brain images. The use of a LDMs allowed increasing the image resolution from 64x64x64 to 160x224x160 without requiring more GPU memory usage or overall training time. More about latent diffusion will be explained in section 3.3.3. To assess the performance of the model and the quality of the synthetic images two main metrics were computed, the Fréchet Inception Distance (FID) for fidelity, and the MS-SSIM for diversity. In both cases DMs metrics ourperformed GANs results.

The first implementation of stable diffusion for medical images, the closest to our work, was introduced by (Chambon et al., 2022) who proposed a model for chest X-ray generation. Their model, named *RoentGen*, was able to create visually convincing, diverse chest X-rays, and the output could be controlled by using text prompts with radiology-specific language. Similar to the work of Pinaya et al., the FID and MS-SSIM metrics were computed although no comparison with GAN-based models was made. A key characteristic of this work is the use of pretrained weights coming from the *Hugging Face Hub*. Instead of training from scratch the network, their suggestion was to fine-tune specific parts of the network to adapt to this new domain. This DM fine-tuning approach is called *Dreambooth* and was first introduced by Ruiz et al. (2023).

The only work we found for lesion inpainting using DM was made for brain MRI by Rouzrokh et al. (2022) from Mayo Clinic. They developed a DDPM to execute several inpainting tasks, like generating lesions or healthy tissue, on slices of the 3D volumes in various sequences. Their model was capable of generate realistic tumoral lesions and tumor-free brain tissue, although the performance of the model was only assessed visually.

2.2. Generative models for mammography

Despite the absence of DM-based model for mammography, other GMs, specially GANs, have been used for several tasks. Wu et al. (2018) tried to tackle the data scarcity and unbalance problem by using class-conditional GANs to synthetically augment mammogram datasets. They focused on training a model for contextual in-filling to synthetize lesions onto healthy screening mammograms. Then, they used this model to generate synthetic images to improve the AUC of a ResNet50 lesion classifier from 0.887 with traditional augmentation to 0.896 with GAN-generated data augmentation.

A full-field digital mammogram (FFDM) generation approach was performed by Korkinof et al. (2019) using GANs as well. Special attention was given to stabilize the GAN training by using stabilization methods and progressive training. The dataset consisted of around 450K images in both MLO and CC view, and they trained the network with a final image size of $1280x1024$. The final size used represented a training stability problem even after using the stabilization techniques. The generated images were only conditioned on the mammogram view. The training was done using 8 V100 GPUs of 16 GB each training for about 70 hours of total training. The trained model was able to generate highly realistic, high resolution synthetic images in appearance, although no quantitative assessment of the fidelity, diversity and radiologist opinion was conducted.

3. Material and methods

3.1. Datasets

We decided to use two datasets for the training of the stable diffusion models (SDMs) to consider different patient populations and mammography unit vendors.

3.1.1. OMI-DB

We used a subset of the OPTIMAM Mammography Image Database (OMI-DB), consisting of around 140k images from several UK breast screening centers (Halling-Brown et al., 2021), various scanner manufacturers and with different image views. The dataset is composed of images with and without lesions (benign, malignant and interval-cancers), and expert annotations are included in the respective cases. Most of the images are available in both *raw* and in *for presentation* format, giving us a total number of $77,035$ images suitable for our generation purpose, distributed among $5,982$ patients. The images were available in DICOM files and its respective metadata was provided in a JSON file format.

Given that the dataset included images from several protocols, such as screening, biopsy and lesion magnification, an extensive image filtering was conducted. The criteria used for the selection of images were the following:

1. No special imaging protocol allowed. E.g. magnification and biopsy images. ($\sim 27k$ images removed)
2. No breast implants ($\sim 3K$ images removed)
3. Only CC and MLO view positions (~ 500 images removed)
4. Only images coming from the *Hologic* mammography units. ($\sim 8K$ images removed)

From the criteria above, only criterion 1 is unavoidable, as we strictly require FFDM. Criteria 2-4 were set to keep a diverse yet uniform data distribution which would be easier for a GM to learn. The generation of minority cases as breast implant mammograms or uncommon mammographic views (such as exaggerated craniocaudal views) were considered out-of-the-scope of this project and let for future work. Because only Hologic images were used the dataset subset is called OMI-H.

The OMI-DB dataset includes metadata at the patient and image level in JSON file format. This information can be accessed using a Python library specifically designed for this dataset. The metadata includes basic DICOM information (laterality, view, pixel size, etc.) and clinical information such as patient status, lesion opinion, biopsy results, and more.

3.1.2. VinDr-Mammo

A second dataset called VinDr-Mammo composed of FFDM with breast-level assessment and extensive lesion annotation was also used. It consists of 5,000 mammography exams, each with 4 standard views (CC and MLO for both lateralities), coming from two primary hospitals from Vietnam, giving a total of 20,000 images in DICOM files (Nguyen et al., 2023). Metadata of each image consisting of both technical and clinical information were also available in a CSV file.

In this case, the only image filtering step performed was to keep images coming exclusively from SIEMENS scanners to avoid learning very different data distributions.

Table 1: Distribution of cases for both datasets.

	OMI-H	VinDr	Combined
Healthy	33,643	13,942	47,585
With lesion	6,908	1,533	8,441
Total	40,551	15,475	56,026

3.2. Data preparation and preprocessing

Both datasets were subject to similar preparation and preprocessing steps. First, all images were saved to PNG format to ensure faster access and less memory usage. Secondly, given the DM architecture used (described in section 3.3.4) the images were saved in RGB format, repeating for each RGB channels the single-channel mammograms, resulting in a visually gray-level image. The original image intensities saved with *uint16* datatype were scaled to a [0, 255] range with a reduced *uint8* datatype.

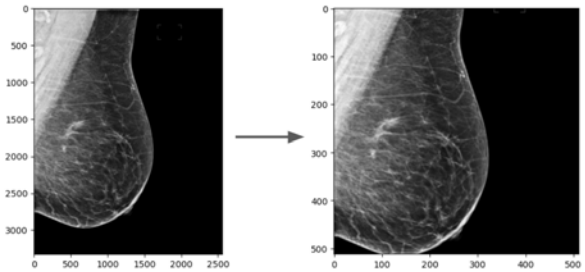
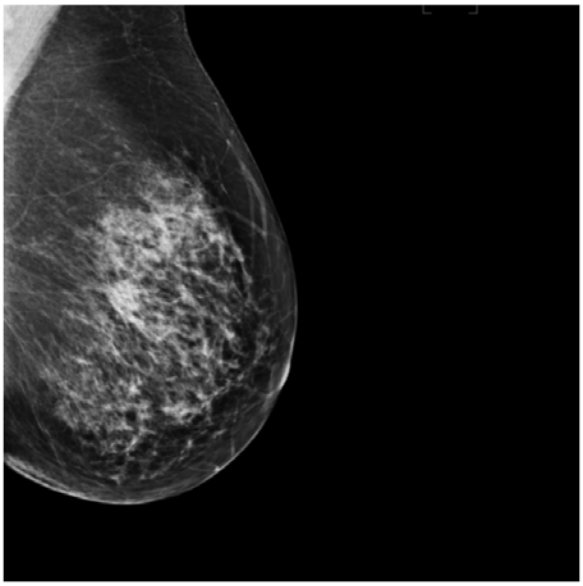
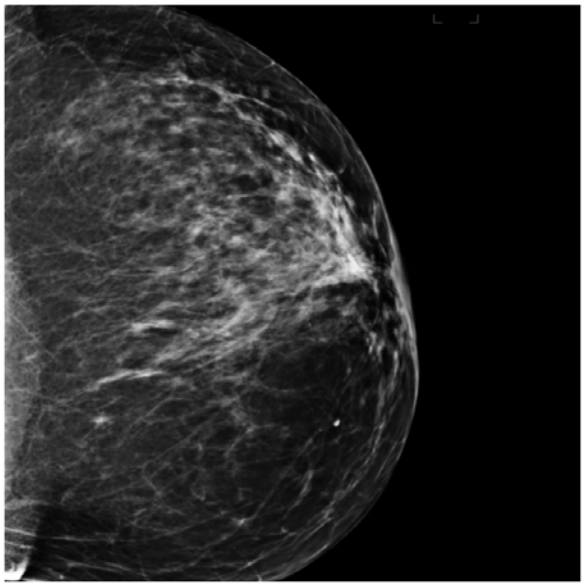


Figure 2: Resizing and cropping of an OMI-H mammogram. The same process was conducted for VinDr mammograms.

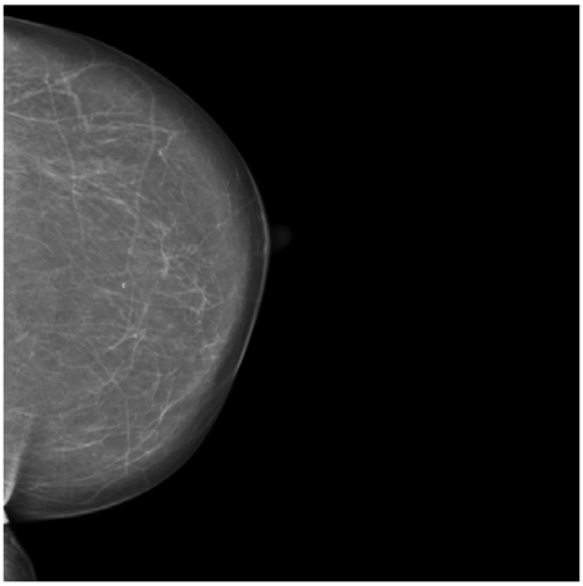
Additionally, in order to use the pretrained weights available for SD, the images were resized to a 512x512 square using bilineal interpolation and center cropping as shown in figure 2. Finally images with right laterality (R) were horizontally flipped so all images have the breast region in the same side, which can potentially facilitate the learning of the data distribution.



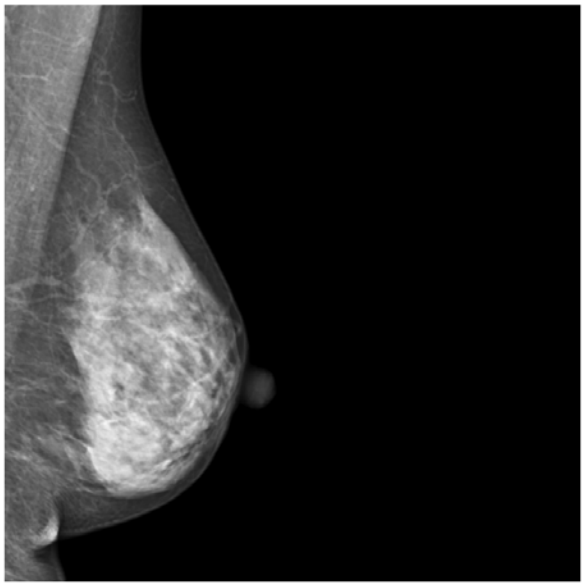
(a) "a mammogram in MLO view with small area"



(b) "a mammogram in CC view with big area"



(c) "a mammogram in CC view with very low density"



(d) "a mammogram in MLO view with very high density"

Figure 3: Examples of training mammograms (real) and their respective text prompts for OMI-H (a-b) and VinDr (c-d).

3.2.1. Task one: healthy image generation

For the first task the healthy images were saved in separated directories, one for each dataset. A text prompt with the description of the image was created and saved along with the image ID in a JSON file. In the case of the OMI-H dataset we created a prompt with the image view and breast area size information. Examples of prompts and their corresponding images are shown in figures 3a and 3b.

To compute the breast area size we first obtained a breast mask using the intensity information of the image and then applying a threshold to separate background and breast tissue. After getting the breast mask we computed the ratio of pixels in the mask compared with the total image. Finally, we define a criteria for three different breast area sizes which can be found in table 2a.

Table 2: Criteria for breast area size and breast density.

(a) Pixel ratio prompt assignment.

Breast area size	
Small	ratio <0.4
Medium	0.4 <ratio <0.6
Big	ratio >0.6

(b) BI-RADS breast density.

Breast density	
Very low	Density A
Low	Density B
High	Density C
Very high	Density D

For the VinDr dataset, we decided not to compute the breast area and, instead, included the breast density information for the prompt description. Breast density was available in BI-RADS scale so we needed to transform this information in a semantically easier text value. We classified the density BI-RADS following the criteria in table 2b. Examples of some images and their prompts can be found in figure 3c and 3d.

3.2.2. Task 2: Lesion inpainting

The second task requires mammograms with confirmed lesions only. Consequently we stored the selected mammograms in separated directories, one for each dataset. Then, using the bounding boxes coordinates available in the metadata, binary masks were generated. Naturally, due to the resizing and cropping preprocessing performed previously, the original coordinates required a proper redefinition using simple geometrical properties.

The binary mask has a pixel value of 255 inside of the bounding box and zero elsewhere. Figure 4 show an example of a mask overlapping a OMI-H mammogram. Because the SD architecture used for the inpainting task requires an input text prompt for the generation, a toy prompt with "a mammogram with a lesion" text was used for all training images.

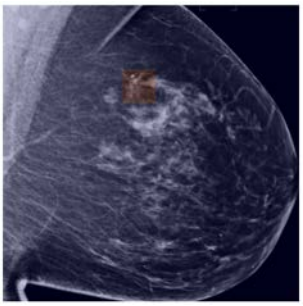


Figure 4: A OMI-H mammogram with a lesion overlapping with its corresponding bounding box mask.

3.3. Diffusion models

The original diffusion model idea was presented by Sohl-Dickstein et al. (2015) and consisted on using a Markov chain[2] to gradually convert one known distribution (e.g. Gaussian distribution) into another (target distribution). Inspired by non-equilibrium statistical physics, the main idea is to systematically and iteratively destroy structure in a data distribution through a process called **forward diffusion**. Then, the **reverse diffusion process** is learned and used to restore structure in data, creating therefore a generative model that implicitly has learned the data distribution.

[2]Defined as a sequence of stochastic events whose time steps depend on the previous one.

The first practical implementation of the DM premise on images was developed by Ho et al. (2020) introducing *Denoising diffusion probabilistic models* (DDPM). In this framework, the data is destroyed by adding Gaussian noise to the image in an iterative fashion described by the Markov chain as shown in figure 5. The total number of diffusion timesteps T is defined by the user but initial experiments were performed with $T = 1000$. To learn the reverse process a encoder-decoder-like neural network (such as a UNet) is used to carry on the denoising process.

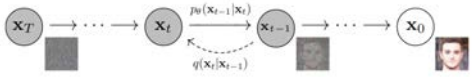


Figure 5: Markov chain of the forward and reverse diffusion process.

A vanilla DM has three main components:

1. Noise scheduler: to add noise in the forward process.
2. UNet: to denoise in the reverse process.
3. Timestep encoder: to encode the timestep t.

3.3.1. Forward diffusion process

Let x_0 be the original image and x_t the noisy version of that image at time t. For the forward diffusion we can define the Markov chain process as

$$q(x_{1:T}|x_0) = \Pi_{t=1}^{T} q(x_t|x_{t-1}), \tag{1}$$

where q is a probability distribution from which the noisy version of the image at time t can be sampled, given x_{t-1}. The proposal of the DDPM framework is to define q as a Gaussian (normal) distribution given by

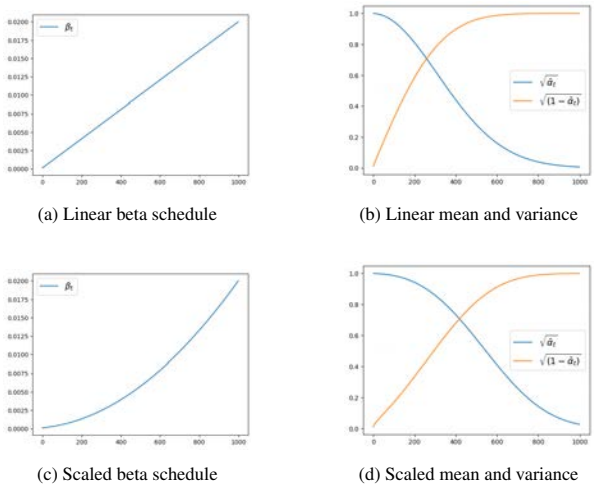
$$q(x_t|x_{t-1}) = N(x_t;\ \sqrt{1-\beta_t}x_{t-1}, \beta_t I), \tag{2}$$

where x_t is the output of the distribution sampling, $\sqrt{1-\beta_t}x_{t-1}$ is the mean and β the variance of the distribution. Therefore the sampling of the next noisy version of the image is essentially controlled by β, as its value affects both the mean and the variance of the sampling distribution. Selecting the manner in which β changes through time is called beta scheduling and is control by the **noise scheduler**. In figures 6a and 6c two examples of beta scheduling are shown.

Thanks to the additive properties of Gaussian distributions, we can obtain a noisy image at any timestep t directly by rewriting the sampling distribution 2 as

$$q(x_t|x_0) = N(x_t;\ \sqrt{\tilde{\alpha}_t}x_0, (1-\tilde{\alpha})I), \tag{3}$$

with $\tilde{\alpha}_t = \Pi_{s=1}^{t} a_s$ and $\alpha_t = 1 - \beta_t$, where α can be interpreted as measure of much information from the previous image is being kept during the diffusion process. The importance of $\tilde{\alpha}_t$, and therefore of β_t, can be understand by looking at figure 6. For t values close to 0



(a) Linear beta schedule (b) Linear mean and variance

(c) Scaled beta schedule (d) Scaled mean and variance

Figure 6: Linear and scaled beta schedulers (left) and their effects on the mean (blue) and variance (orange) of the noise sampling distributions (right).

the distribution from which we sample have $\mu \approx 1$ and $\sigma \approx 0$, meaning that the sample images are every similar to the original image. On the other hand, for large t values where $\mu \approx 0$ and $\sigma \approx 1$ the distribution is close to a standard normal distribution (SND) and the sampled image will be essentially pure Gaussian noise.

Finally, to be able to define the training goal in the reverse diffusion process, we express the sampling from the probability distribution in equation 3 using the *reparameterization trick* (Kingma and Welling, 2022). The reparameterization trick allows us to write the generation of a sample X from a normal distribution $N(\mu, \sigma)$ as $X = \mu + \sigma Z$, where $Z \sim N(0, 1)$, i.d. Z was sampled from a SND. With this, the forward diffusion sampling process can be expressed by

$$x_t = \sqrt{\tilde{\alpha}_t} x_0 + \sqrt{1 - \tilde{\alpha}_t}\epsilon, \tag{4}$$

where $\epsilon \sim N(0, 1)$. The stochastic variable epsilon (ϵ) in equation 4 is crucial to understand the reverse diffusion process as it is basically the prediction target of the UNet.

3.3.2. Reverse diffusion process

The reconstruction of the data destroyed by noise can be done using a UNet that has learned to denoise the images. Formally, the reverse process is also a Markov chain that can be defined in a similar way as

$$p_\theta(x_{0:T}) = p(x_T)\Pi_{t=1}^{T} p_\theta(x_{t-1}|x_t), \tag{5}$$

where p_θ is the learned probability distribution from which the denoised images are sampled at each timestep t. θ indicates that the distribution is parameterized as it was learned by the UNet. This also explains why the term $p(x_T)$ has no subscript θ as it is the starting point of the reverse process, i.e. pure Gaussian noise.

Assuming that p can also be modeled as a normal distribution, it can be expressed as

$$p_\theta(x_{t\_1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \tag{6}$$

where μ_θ and Σ_θ are the learnable mean and variance of the reverse sampling distribution. To reduce the training complexity, and because it showed to give similar results, $\Sigma_\theta = \beta I$, therefore only μ_θ has to be learned. Sadly, due to limitations of space in this report, the complete formulation of the optimization of the usual variational bound on negative log likelihood cannot be fully described. Key considerations of this formulation are given instead.

The first consideration is that μ_θ can be computed as

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_T}}(x_t - \frac{\beta}{\sqrt{1 - \tilde{\alpha}_t}}\epsilon_\theta(x_t, t)), \tag{7}$$

where the key is to notice that we only need to predict ϵ_θ to predict μ_θ.

The second consideration is that the term we need to optimize, and which consequently defines the loss function of our UNet, is

$$L = \|\epsilon - \epsilon_\theta\|^2, \tag{8}$$

where epsilon (ϵ) is the same Gaussian noise we defined in equation 4, sampled from a Gaussian distribution $\epsilon \sim N(0, 1)$, and ϵ_θ is the output of the UNet. In other words, the UNet objective is to implicitly learn the data distribution by predicting the scaled Gaussian noise ϵ added to the images at timestep t.

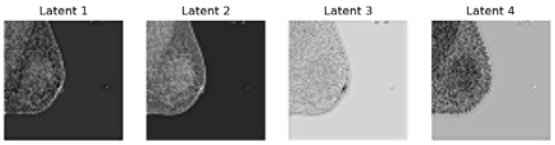
Finally, to include the timestep as an additional input to the Unet, a timestep encoder is used to embed this information and use it during training. More information will be given in section 3.3.4.

3.3.3. Latent diffusion

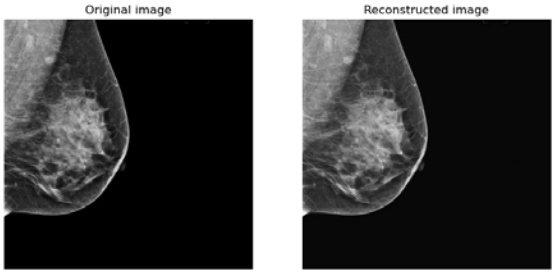
Image size is one of the main constrains when training generative models. Medical images usually require high resolution, specially in the case of mammograms where the sizes go up to 3 or 4 thousand pixels per image side. Training a DM for such sizes would require large computational resources and extensive training time.

Latent diffusion tries to solve this issue by using encoders to *compress* images from their original sizes in the image space into a smaller representation in the latent space. The motivation behind this is that images usually contain redundant information and an encoder can produce a smaller representation that can later be reconstructed back using a decoder.

An example of this is shown in figure 7, where an original image of 512x512 pixels is compressed to 4 latent representations of 64x64 using a Variational Autoencoder (VAE), reducing 16 times its original shape (Kingma and Welling, 2022).



(a) Latent representations.



(b) The original image and its reconstruction.

Figure 7: Example of the latent space representation of an image and its reconstruction.

Consequently, as introduced by Rombach et al. (2022), in latent diffusion the diffusion processed (described in the previous section) is performed on the latent representations rather than the original images. This allows using diffusion pipelines with lower memory usage, fewer layers in the UNet, and faster training and generation.

There exist different types of encoders that can be used and the selection criteria reside mainly on the type of images and their task. Chambon et al. (2022) found that a pretrained VAE on natural images can have good perform in medical images as well. For this reason we decided to use a VAE for our work, obtaining visually successful encoding for mammograms as shown in figure 7b. More information on the model architecture can be found in section 3.4.

3.3.4. Stable diffusion

Pure latent diffusion does not include conditioning at training or inference time, and synthetic images are generated from the learned distribution, depending on the starting Gaussian noise. Stable diffusion is an improvement to Rombach et al. (2022) work, in which text conditioning is added to the model for additional control on the generation process.

In stable diffusion the text conditioning is a prompt with the description of the image. To create a numeric representation of the prompt we use a pretrained transformer called CLIP (Radford et al., 2021). CLIP, which stands for Contrastive Language-Image Pre-training, maps both text and images into the same representational space, allowing comparison and similarity quantification between them (Frans et al., 2021). In other words, CLIP allows us to compare images and text.

CLIP first uses a subword-based tokenizer to convert any prompt text to a fixed 77 tokens length. Then, the CLIP encoder sends each token into a 768-dimensional vector, which lives in the image-text CLIP space. The CLIP embedded text is then used in the attention layers of the UNet through a cross-attention mechanism. More details are given in section 3.4.

3.3.5. Fine-tuning SD: DreamBooth

In 2022 Stability AI and LAION made the pretrained weights of Rombach et al. (2022) model publicly available, which allowed the GM community to train domain-specific fine-tuned SD models. Nevertheless, fine-tuning a large text-to-image model and teaching it new concepts can be challenging and one can face several difficulties such as catastrophic forgetting[3], overfitting and low image generation diversity.

Ruiz et al. (2023) presented an approach for fine-tuning SD called *DreamBooth*. They proposed using only a few images of the new subject with its respective text prompt, to train the model using a small learning rate. Additionally, if the subject semantically exists in the model domain, prior generation images can be included for the training. This allows the binding of the new subject to a new unique identifier in the text embedding space, as well as a learned representation in the pretrained data distribution.

This fine-tuning technique has been tried for chest X-ray by Chambon et al. (2022) and showed promising results on adapting the SD domain into their images to generate high-fidelity and diverse images thanks to the control given by the text prompt.

3.4. Task 1: Normal mammogram synthesis

We propose putting together all the pieces presented above and adapting the Dreambooth fine-tuning technique for mammograpic images generation, using the pretrained *stable-diffusion-v1-5* model as baseline, publicly available in the *Hugging Face* model hub repository.

For each dataset we decided to train a separate model using only healthy images, as each dataset contains independent semantic information in the prompt and because the intensity ranges and image details differ between populations. This means that we trained separate models for *Siemens* and *Hologic* mammograms. Additionally, we decided to train a combined model with images of both vendors, adding in the prompt text the vendor's name.

Our SD pipeline has three independent models that could potentially be trained at the same time. Given the good performance of the VAE encoder on mammograms, we decided to keep it frozen an train only the CLIP text encoder and the UNet weights. The three models are summarized as follows:

[3]This means the model forgetting previous information and concepts.

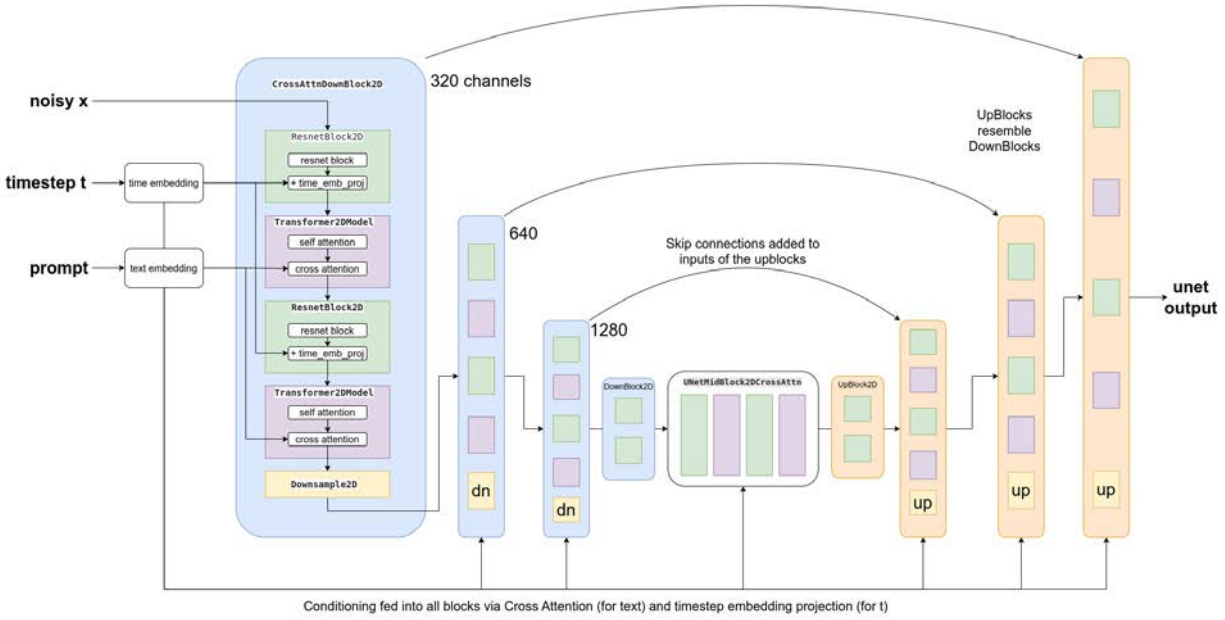


Figure 8: Training pipeline for full mammogram generation. UNet architecture from the original stable diffusion paper (Rombach et al., 2022).

1. VAE as image encoder (frozen): Pretrained with LAION-b5
2. CLIP text encoder: *CLIP ViT large* using a ViT-L/14 Transformer architecture from Open AI.
3. UNet: Pretrained with LAION-b5.

The pretrained VAE model inputs RGB 512x512 and outputs latent representations of 64x64x4, just as shown in figure 7.

3.4.1. Training

The UNet architecture is the original SD UNet proposed by Rombach et al. (2022) and its presented in figure 8. The network has 4 2D down- and upsampling blocks. Except for the last downsampling block (and its corresponding upsampling block) all blocks are composed of two ResNet blocks and two Transformer blocks, one after the other. The timestep embedding is added to the ResNet blocks whereas the text embedding is added through cross attention into the Transformer blocks. For the last downblock (and first upblock) only the timestep information is fed.

One training step can be summarized as follows:

1. Sample a batch of images $x_0 \sim q(x_0)$
2. Encode x_0 into the latent space
3. Sample a random timestep from a uniform distribution $t \sim U(1, ..., T)$
4. Sample random Gaussian noise from a normal distribution $\epsilon \sim N(0, I)$
5. Create x_t by adding noise to the batch images x_0 using the noise ϵ and timestep t.
6. Take a optimization step in the direction of the gradient $\nabla_\theta \|\epsilon - \epsilon_\theta(x_t, t)\|$
7. Repeat

We notice that, contrary to what is popularly believed, the DM training process does not consist on denoising the same image in a sequential order. This confusion comes from the way the diffusion process is presented as a Markov chain in figure 5. Instead, the training encompasses three stochastic processes by randomly sampling the main components of the diffusion process: the original image x_0, the Gaussian noise ϵ and the timestep t. By doing so we avoid the overfitting of the network on the sequential way the images are given and focuses on the denosing process per se.

The CLIP text encoder and the UNet parameters are updated simultaneously at each training step using an AdamW optimizer, with the MSE as loss function described in equation 8.

The main training hyperparameters (HP) are the following:

- Batch size: We ranged from 8, 16, 32, 64, 128 and 256.
- Training steps: Experiments ranged from 1k up to 16k.
- Learning rate: We explore three main values $1e^{-6}$, $1e^{-5}$, $1e^{-4}$.

Other HP that were not changed include: constant lr schedule, Adam weight decay and epsilon, gradient clipping and dropping the last incomplete batch per epoch. All detailed HP information can be found in the configuration file of each experiment in the GitHub repository.

We generated 4 sample images every 100 or 200 training steps to track the performance of the models, as well as the training loss. This was loaded to the cloud using *Weights and Biases* (WandB) logging tool. Additionally all models were uploaded to the authors personal Hugging Face repository and are publicly available.

3.4.2. Inference: image generation

With the UNet prediction we can denoise pure Gaussian noise and generate new mammograms. The procedure is as follow:

1. Sample random Gaussian noise $\epsilon \sim N(0, I)$
2. for $t = T, ...1$ do:
3. $z \sim N(0, I)$ if $t > 1$ else $z = 0$
4. $x_{t-1} = \frac{1}{\sqrt{\alpha_T}}(x_t - \frac{\beta}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)) + \sigma_t z$
5. end for
6. Decode image using VAE

First, random Gaussian noise is sampled as starting point. Then the denoising process is repeated for T steps. The loop consists on using the predicted noise ϵ_θ to compute the distribution mean using equation 7. By adding $\sigma_t z$ to this mean term we are essentially sampling from the learned data distribution of the reverse diffusion process. After the denoising process is finished the image is send back to the image space using the VAE decoder.

The inference process has two main HP to consider: number of timsteps T and the guidance scale. First, the number of timesteps T will depend on the type of sampling method that we use for denosing. The traditional DDPM sampling requires around 100 steps to generate good quality images, which is time consuming and represent a bottleneck in the image generation. The best alternative we found is to use the DPM-solver proposed by Lu et al. (2022), which allows fast diffusion sampling with only 20 steps for good quality image generation. In the result section we show how the change of T affects the image quality.

The second HP is called the guidance scale. Even though the SD architecture uses cross attention in several parts of the network, so the generation process focuses on the text prompt, in reality this is still not enough and the model tends to ignore the text prompt at inference time. To solve this issue Ho and Salimans (2022) proposed a technique called classifier-free guidance.

In essence, classifier-free guidance consists on generating two noise predictions ϵ at each step, one using the prompt (ϵ_{text}) and one without it (ϵ_{free}). Then, the difference between the prompt-generated noise and the free-generated noise is computed. This difference can be considered as a vector in the image distribution space, which points in the direction of the image with text. As such, we can scale this vector and sum it to the free-generated noise to force it to go more in the direction of the prompt text. This geometrical trick is illustrated in figure 9.

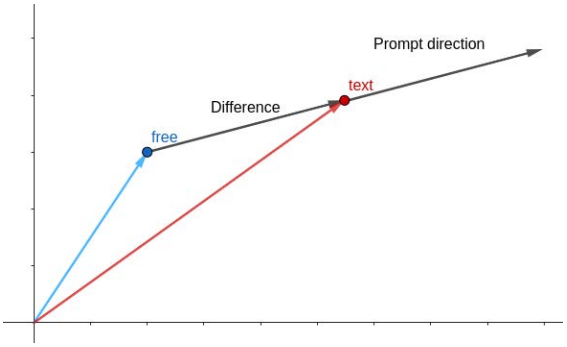


Figure 9: Classifier-free guidance geometrical interpretation. As the guidance scale increases, the image is pushed further in the prompt direction.

Formally, the scaling factor is called guidance scale and the formulation can be summarized as follows:

$$\epsilon_\theta = \epsilon_{free} + guidance * (\epsilon_{text} - \epsilon_{free}). \quad (9)$$

3.5. Task 2: mammographic lesion inpainting

The SD pipeline described for task 1 can be modified in some key aspects to be able to perform the inpainting task. We propose using the modified DreamBooth fine-tuning pipeline to inpaint lesion in a designated region of the breast. To the knowledge of the authors, this is the first work to use SD fine-tuning for lesion inpainting in medical images.

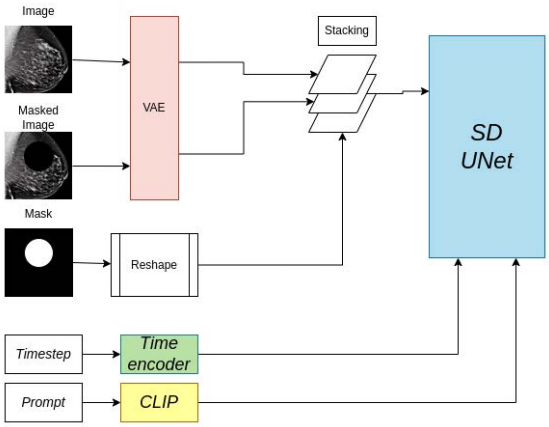


Figure 10: Inpainting training pipeline. The same UNet as in the SD pipeline in figure 8 is used.

At the dataloader level, for each batch two new elements are added per example: the mask and a masked version of the original image. The masked version means that the pixel values inside of the bounding box are set to zero.

At training time, first both the image and the masked image are encoded using the latent space. Also the mask

must be reshaped to the latent representation size. The rest of the diffusion process remains the same, only one crucial difference is made: instead of feeding the latent representation only to the UNet, the latent representation, the mask, and the masked latent representation are stacked into one tensor. This new input is then fed to the UNet, as well as the encoded timestep and prompt text as in a traditional SD. This process is described in figure 10.

This small change in the training process allows the network to pay attention only to the pixels inside the mask, as the pixel outside of it are always provided. The rest of the pipeline follows the same principles as the ones we described for task one.

3.6. Complete MAM-E pipeline

The models used in task 1 and task 2 can be put together in a sequential order so that a full synthetic mammogram with lesion can be generated. Figure 11

3.7. Resources management

Having three large models loaded at the same time, and enabling the gradient tracking for two of them for training, can represent a dramatic increase of GPU and processor resources. Thankfully there exist a handful of techniques and frameworks to reduce this demand and fit the training in a GPU memory of circa 20 GB with an efficient batch size of 256.

First, we used mixed precision using the *fp16* arithmetic, and the revision model (model version) specifically for that precision. When training in the (Ampere) A30 or A40 GPU we activate the bf16 precision, with no improvement in the time or apparent quality of the training.

We also used lighter version of the AdamW optimizer, the 8-bit AdamW optimizer by *Bitsnadbytes*. Additionally, because our three models use attention layers, we made us of the *Xformers* efficient memory usage for transformers which speeds the training time and decreases the GPU usage.

To achieve the 256 batch size in one single GPU we used gradient accumulation, a technique that consists on computing the gradient for a mini-batch without updating the model variables, for a set number of times, summing the gradients. By doing so, the general batch size is accumulated an essentially the batch size increases. In our case, using a mini-batch size of 16, and 16 gradient accumulation steps the accumulated batch size is 256. This technique, although clever comes with an increase in training time.

Gradient checkpointing is another technique to use the CPU processors power to help release some GPU memory at the expenses of a slower training. Gradient checkpointing saves strategically selected activations throughout the computational graph of the model tensors so only a fraction of the activations need to be re-computed for the gradients.

Finally one can simply set the optimizer gradients to None instead of zero after the weights update have been completed. This will in general have lower memory footprint, and can modestly improve performance.

Most of these techniques can be implemented directly using *Hugging Face Accelerate* library and framework for distributed training and resources management.

3.8. Assessment

The assessment of generative models depends on the application of the synthetic images and it may not be straightforward as in other DL models. While it is possible to observe the changes in the loss values during training, the loss curve rapidly converges to a specific region and no further difference is notice. It is specially difficult to see any substantial loss differences when performing DreamBooth fine-tuning.

Regardless of the apparent plateaued loss function, the semantics learned by the model are continuously changing during training. One possible way to keep track of the model training performance is to log examples of synthetic images every several steps to see this semantic changes.

At inference time there exists other types of assessment and they can be categorized as follows:

- Qualitative assessment: Focusing on the visual appearance of the mammograms.
- Quantitative assessment: Computing metrics to attest the diversity and fidelity of the generated images, as well as generation time.
- Quantitative CAD assessment: Exploring the potential benefits of synthetic images on CAD systems performance.

3.8.1. Qualitative assessment

We performed two types of visual assessment. First, an overall simple visualization of the images to see any clear inconsistency (noise remnants, anatomical irregularities).

Then, we assessed the quality of the images by asking a radiologist with 30 years of experience to rate 53 mammograms, with a 50/50 real-synthetic ratio, in a scale from 0 to 4, using the following criteria:

- 0: Definitely real
- 1: Probably real
- 2: Not sure
- 3: Probably synthetic
- 4: Definitely synthetic.

These results were then converted to probabilities to be able to obtain a ROC curve and its respective area. The main objective is to attest the radiologist's ability to differentiate synthetic mammograms apart from real ones.

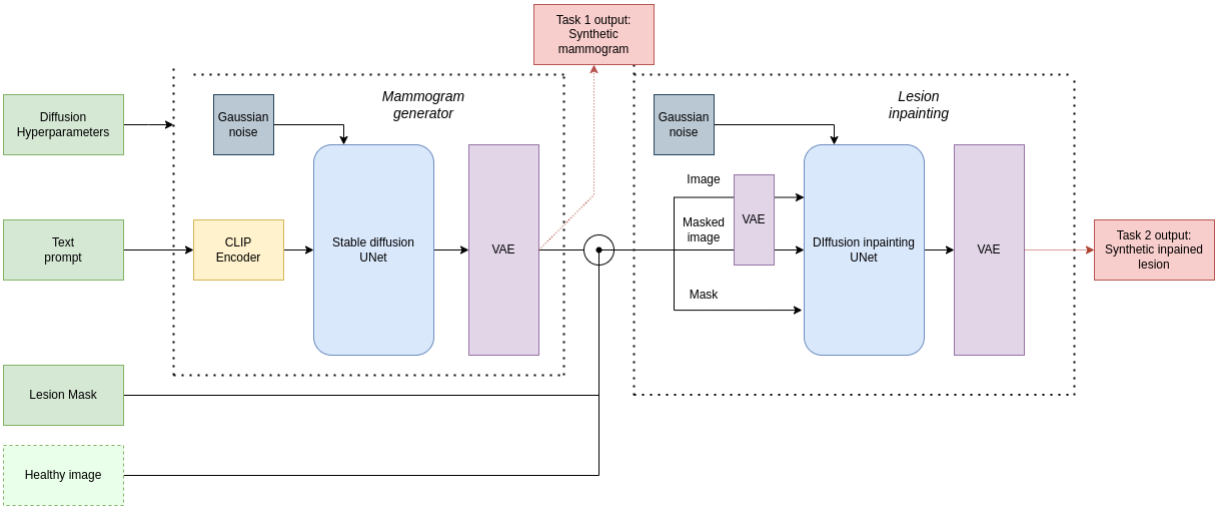


Figure 11: The complete MAM-E pipeline combining task 1 and task 2 pipelines. In dark green, the inputs needed for a full synthetic mammogram generation with lesion. In light green the optional input for lesion inpaining on real images, overriding task 1. In red, the outputs of each task.

3.8.2. Quantitative assessment

Generative models are expected to have two main characteristics: generation diversity and fidelity to the original dataset. There exists metrics to quantitatively assess them.

Generation diversity can be computed using the pairwise Multi-scale structural similarity index metric (MS-SSIM). If a pair of synthetic images are sampled, a low MS-SSIM value would mean that the compared images are not structurally similar and, therefore, implies diversity.

Fidelity can be calculated using the Fréchet Inception Distance (FID), a metric first proposed for GAN-generated image quality assessment by Heusel et al. (2018). FID captures the similarity of generated images to real ones by comparing the statistics of a collection of synthetic images and a collection of real images.

Formally, the activation vector from the last pooling layer of a ImageNet-pretrained Inception V3 is computed for a set of real and synthetic images. This 2048-vector is called the feature vector and contains computer-vision-specific information of the images. The FID consists, then, on calculating the Fréchet distance between the Multivariate Gaussian distribution of both population, synthetic and real. This is done by sampling N examples from each distribution. N is recommended to be $10,000$ for the best approximation, although some works suggest using a lower N number can also be representative of the distribution.

Notwithstanding, the use of the FID metric is controversial and even discoraged. Chong and Forsyth (2020) found that the FID is biased towards the generative model and should not be used. Moreover, given that no other work has explored the FID for FFDM generation we decided not to compute the FID and assess the image fidelity based on the radiological expert visual assessment of the previous section.

Finally, generation time has to be assessed as the denoising time of DM is one of its main drawbacks. Generation time is closely tied to other inference HP like the guidance scale or the prompt length and order.

3.8.3. Quantitative CAD assessment

The performance of the generative models and the utility of the images for training CAD models can be assessed using CAD pipelines and referring to the effect of adding synthetic images during training.

We decided to collaborate with a master thesis defendant, Sam-Millan (2023) from the ViCOROB lab and whose works are included in the proceedings of this year.

Sam-Millan et al. worked on explainability AI (EAI) for patch classification and full-field mammogram lesion classification. Specifically for FF mammogram classification problem, the EAI system explores which regions of the image are more relevant for the classification task. A heatmap of these regions importance is generated for several EAI methods. It is expected that the lesion region captures the main attention.

We asked the authors to create heatmaps of a healthy real mammogram with a synthetic inpainted lesion to assess if the classifier is focusing on the synthetic lesion region.

For more details on any of these methods refer to the corresponding report.

4. Results

4.1. Training unconditional model

The first experiments were conducted on unconditional diffusion models, meaning that no text prompt guidance was used as input. This first steps were vital

for the building of the conditional models later on. Visual assessment of the generated images during training time is presented in this section.

4.1.1. Image space: vanilla DM

The first trial consisted on generating 64x64 mammograms to observe the evolution and behavior of DM while trained from scratch using mammograms.

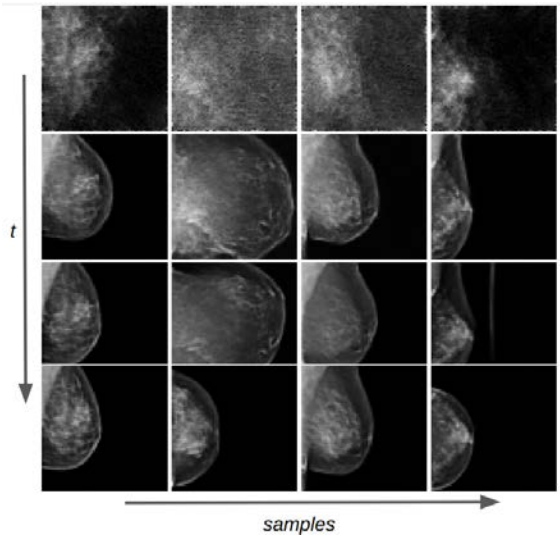


Figure 12: Training evolution of the vanilla image space diffusion model at 1k, 2k, 3k and 4k timesteps. This corresponds to epoch 1, 16, 3 and 50.

The corresponding loss and log-loss function of this vanilla DM are presented in figure 13.

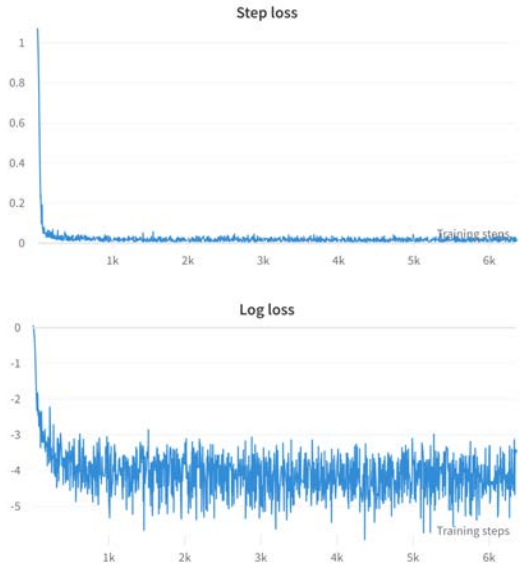


Figure 13: On the top: Loss of the vanilla DM training for 6k steps. On the bottom: the corresponding log loss.

This vanilla DM training helped us to make two initial main considerations. First, it allowed us to assess the training of a diffusion model from scratch and how the denoising process can effectively generate meaningful images after the first 2k training steps, as figure 12 shows.

Secondly, we observe that the loss function rapidly reaches a plateau, which show how fast the loss objective can be minimized in diffusion. Nevertheless, as seen in the log-loss, the function fluctuates in a specific range. This is a common behavior of DM losses, which tend to reach a stability region where the loss varies and then slowly starts to decrease as the denoising process is learned.

4.1.2. Latent space diffusion

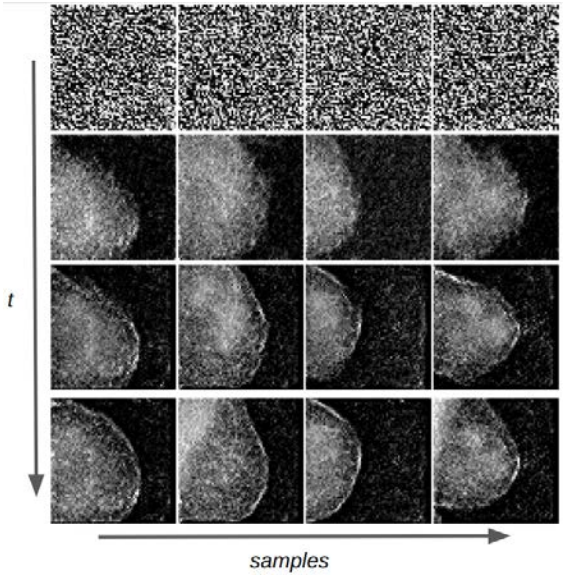


Figure 14: Training evolution of only one latent representation at epoch 1, 16, 36 and 50.

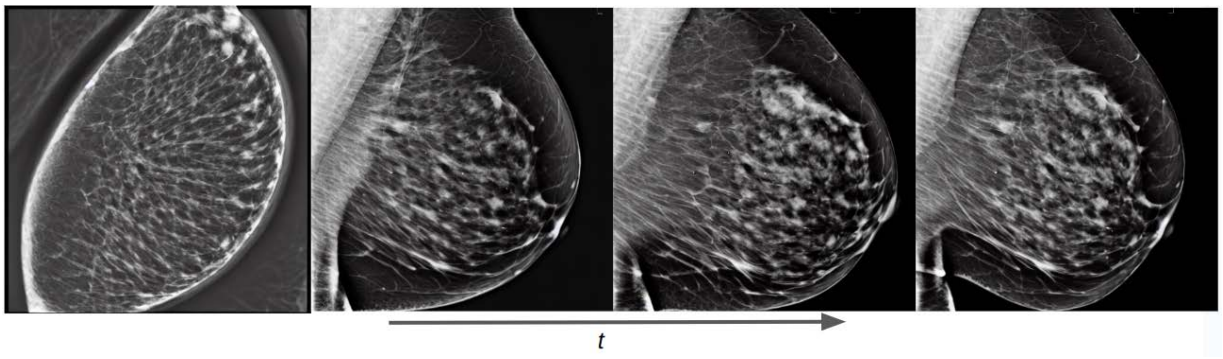
The second group of experiments consisted on sending the images to the latent space and implement the diffusion process on the latent representations. Figure 14 shows the latent denoising process for one channel (there are 4) of the latent representation.

It can be seen that the denoising process is more difficult and slower in the latent space. After 50 epochs the latent image still presents remnants of the original Gaussian noise. In contrast, in the image space the image present almost no signs of Gaussian noise after the same number of training epochs are completed.

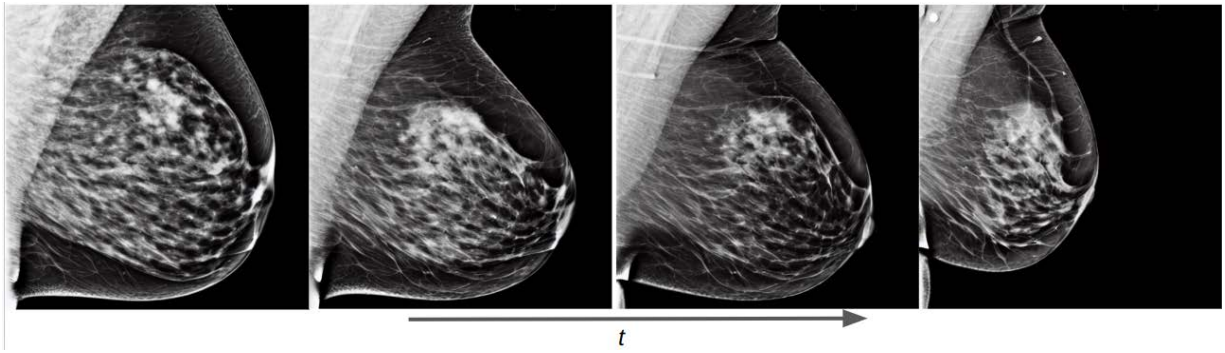
This behavior is expected as the prediction objective of the UNet in the image space consists of predicting a 64x64 one-channel ϵ matrix. On the other hand, the prediction objective in the latent spaces is a 54x54 four-channel ϵ matrix, 4 times bigger representation, which represents a more challenging objective to train.

4.1.3. Unconditional pretrained models

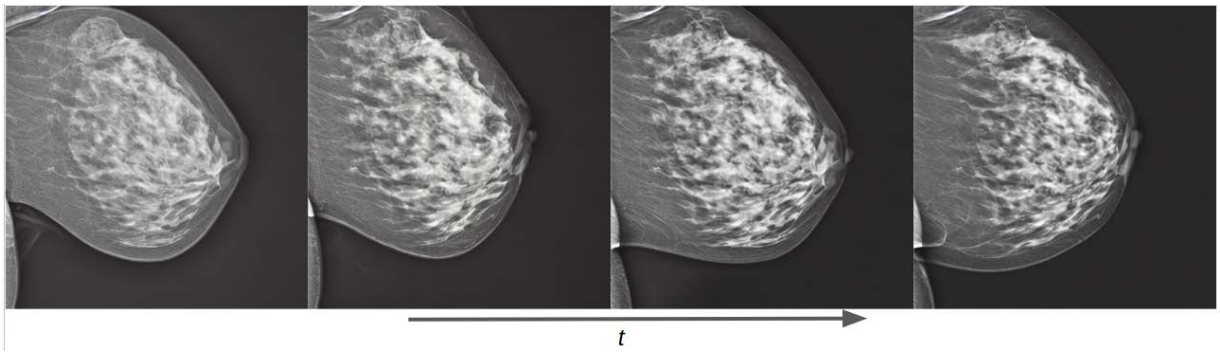
We solved the issue of denoising the mammograms latent representation using a pretrained unconditional



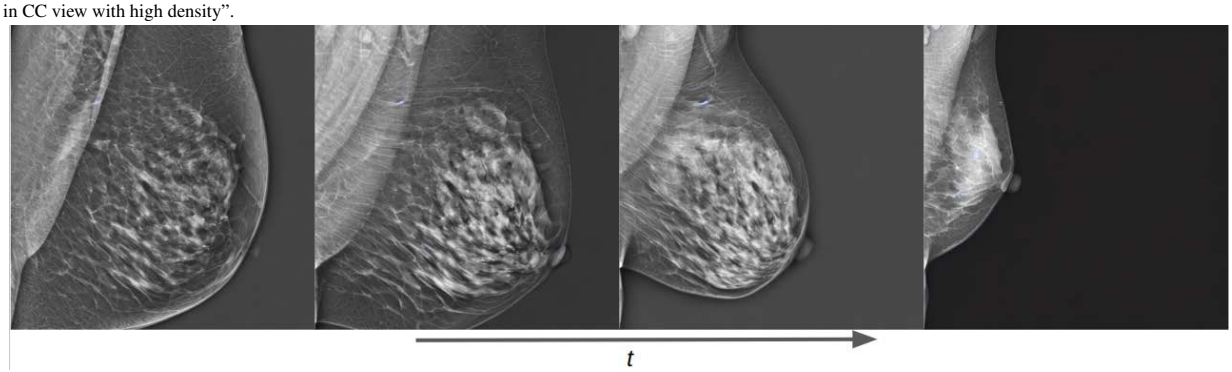
(a) Training evolution of the diffusion process on an unconditional pretrained model at epoch 1, 3, 6 and 10.



(b) Training evolution of the diffusion process on a conditional pretrained model trained with Hologic images at epoch 1, 3, 6 and 10. The prompt is: "a mammogram in MLO view with small area".



(c) Training evolution of the diffusion process on a conditional pretrained model trained with Siemens images at epoch 1, 3, 6 and 10. The prompt is: "a mammogram in CC view with high density".



(d) Training evolution of the diffusion process on a conditional pretrained model trained with both Siemens and Hologic images at epoch 1, 3, 7 and 40. The prompt is: "a siemens mammogram in MLO view with high density and small area".

Figure 15: Training evolution for several diffusion processes.

latent diffusion model and fine-tuning it. Figure 15a shows the evolution of the diffusion process as the training steps progress. It can be seen that from the first epoch the generated image has essentially no signs of residual Gaussian noise, although the synthetic image does not resemble a mammogram. This implies that

the pretrained model has already learned how to denoise images and that the new task is to learn a new concept (a mammogram) and find its representation in the data distribution of the model.

We can also notice that in only 3 epochs the model has already learned the fundamental characteristics of a mammogram and can generate realistic images. In the following epochs the model focuses on improving smaller details on the image, like the edges of the breast and the details of the breast parenchyma.

4.2. Fine-tuning conditional models: DreamBooth

The training of the conditional model using prompt text can be shown in figure 15b for the Hologic dataset and 15c for the Siemens dataset.

First, we observe that the conditional model, besides learning the anatomical structure and form of a mammogram, pushes the generated image in the direction of the text prompt semantics as the training process increases. In the case of the Hologic training, in figure 15b we can see that the mammogram reduces its shape in accordance to the area described in the prompt text.

In the case of the Siemens example in figure 15c, the image view starts in a tiled position similar to MLO but it is slowly corrected to match with the prompt description, that is a CC view. Similarly, the apparent breast density is kept high, in accordance to the input prompt.

Therefore we can acknowledge that a conditional model, thanks to the combined training of the CLIP text encoder and the UNet, learns to modify the generated image to better match the generated pair image-prompt similarly to how they are paired in the training set.

4.3. Fusion MAM-E: combined datasets

The combination of both datasets allowed us to train a model we called **Fusion MAM-E**. Besides allowing us to also select the vendor type of the generated mammogram, this model allowed us to extrapolate the characteristics of one dataset to the other. This means that, e.g. the breast density of the Hologic mammograms could be controlled, even though this information was not available in the Hologic dataset.

Figure 15d show some training samples generated at different epochs. During the first epoch, the model generates a synthetic images with in the correct view but with large breast area and low breast density. We observe how, as the training process advances, the breast area shrinks and the breast density augments, in accordance with the prompt. Also, the image intensities and overall texture starts to be more similar to a Siemens mammogram.

It is also valuable to remark the gray background present during the first training epochs of the fusion model. The complete removal of this feature, shifting it into a black background, required 40 training epochs, a considerable difference with the other conditional models. This is expected as the fusion model incorporates more semantic concepts to the CLIP text encoder that have to be learned, as well as a larger image dataset.

4.4. Quantitative assessment

All synthetic examples above, even though logged during training time, naturally involves a inference pipeline. As explained in section 3.4.2 there are two main HP that have to be tuned during inference.

First, the denosing steps T must be set. In our case, because we used the DPM-solver of Lu et al. (2022) we only needed, in general, 24 timesteps for denoising. This usually means an average time of 2 seconds for the denoising of one sample. In some cases, due to the increase of the guidance scale, the number of T steps must be increased to completely remove the noise. The longest generation samples that we run used $T = 50$, needing maximum 4 seconds to denoise.

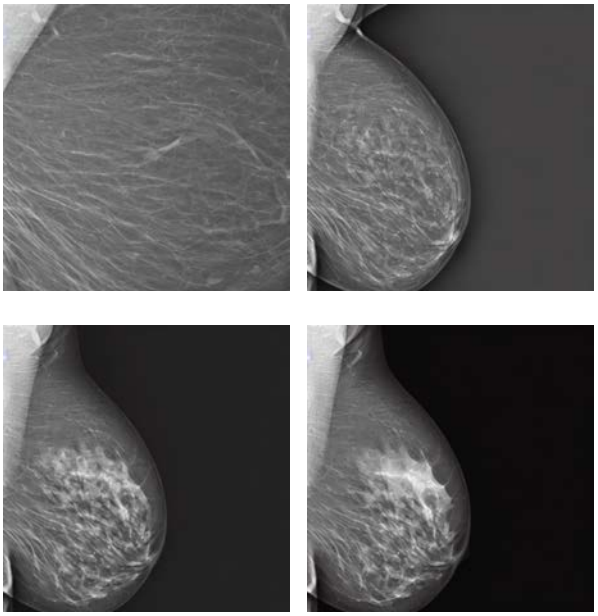


Figure 16: Guidance effect on the generation output. From upper-left to lower-bottom the guidance varies in a range from 1 to 4. Prompt: "A siemens mammogram in MLO view with small area and very high density".

The guidance scale, on the other hand, played a more crucial role in the quality and diversity of the generated images. Figure 16 shows the effect of the guidance scale on the image generation.

First, we observe that a guidance scale of 1 does not suffice for a meaningful generation. This is a common behavior for stable diffusion pipelines, as the image must be pushed further in the prompt direction (figure 9). It can be seen that the increase in the guidance value not only generates a more meaningful image, but also adjusts the characteristics of the mammogram to better match the text prompt. For example, at guidance 2, the mammogram still presents low breast density. In

the following 3 and 4 guidance values the breast density increases, as well as the overall quality of the image.

Nevertheless, there exists a trade-off between prompt fidelity and generation diversity. If the guidance scale is high, the generated images may all look similar, creating some kind of ”mode collapse” for DM.

To quantitatively assess this phenomenon, we computed the MS-SSIM metric for different guidance scale values. The mean and standard deviation of the MS-SSIM value among 20 images of the same prompt and guidance value were computed and are shown in table 3. The experiment was repeated for the two vendors and the fusion model. The prompt was randomly selected for each model.

Table 3: Guidance scale effect on the MS-SSIM of the three SD models. The lower the MS-SSIM the higher the image diversity.

	Hologic		Siemens		Fusion	
Guidance	Mean↓	STD	Mean↓	STD	Mean↓	STD
4	0.29	0.16	0.38	0.19	0.37	0.14
5	0.34	0.16	0.36	0.17	0.44	0.16
6	0.38	0.12	0.41	0.17	0.51	0.15
7	0.38	0.1	0.34	0.17	0.49	0.19
8	0.43	0.11	0.42	0.2	0.53	0.14
9	0.42	0.13	0.43	0.16	0.44	0.17
10	0.49	0.12	0.41	0.13	0.6	0.11
11	0.5	0.12	0.47	0.17	0.51	0.14
12	0.52	0.11	0.46	0.16	0.47	0.12
13	0.48	0.1	0.42	0.16	0.51	0.17
14	0.5	0.11	0.4	0.18	0.47	0.14

From these results, it can be seen that, overall, the higher the guidance value the lower the generation diversity, as the MS-SSIM value decreases. This suggests that the value of the guidance scale must be carefully selected as a very low value will generate low quality images but with high diversity. Conversely, a high guidance value (above 6) will generate a mammogram more faithful to the prompt description but with low diversity.

Also, we attest that the optimum guidance scale will depend on the model, so empirical experiments using the MS-SSIM metric are encouraged.

4.5. Qualitative visual assessment

A more formal visual assessment was performed with the radiological evaluation of 53 synthetic image by a radiologists. The results of the test are summarized as a ROC curve in figure 17. The shape of the ROC curve bears resemblance to the random guess curve, suggesting that the radiologists cannot easily identify the difference between real and synthetic images. Moreover, the AUROC value obtained by the radiologist for this synthetic classification task was 0.49.

4.6. Quantitative CAD assessment

The heatmaps of six Explainability AI methods, computed by Sam et al., were obtained for a healthy mammogram with an inpainted synthetic lesion. All six

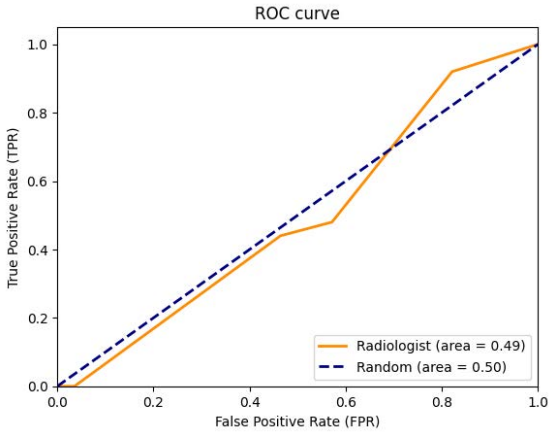
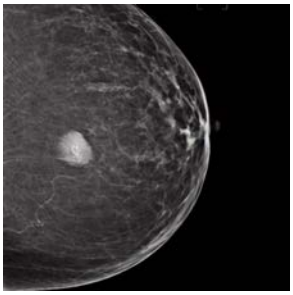
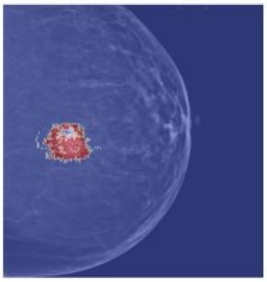


Figure 17: ROC curve of radiological assessment.

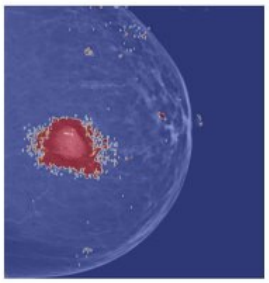
methods can be found in Sam-Millan's thesis report. here we present only three methods: gradcam, saliency and occlusion, with their respective heatmaps in figure 18.



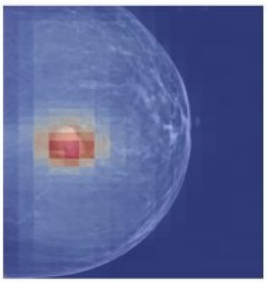
(a) Inpainted synthetic lesion.



(b) Gradcam heatmap.



(c) Saliency heatmap.



(d) Occlusion heatmap.

Figure 18: Explainability AI methods heatmaps of synthetic lesion over real healthy mammogram.

All three maps show that the classification method used by Sam et al. focuses on the synthetic lesion to make the prediction. This means that this CAD system is sensible to the presence of the lesion, which suggests that it may contain a pixel distribution similar to those present in real images.

4.7. MAM-E Graphical user interfaces

We decided to build GUIs to make the pipelines of both tasks available and easy to use to the public. Our

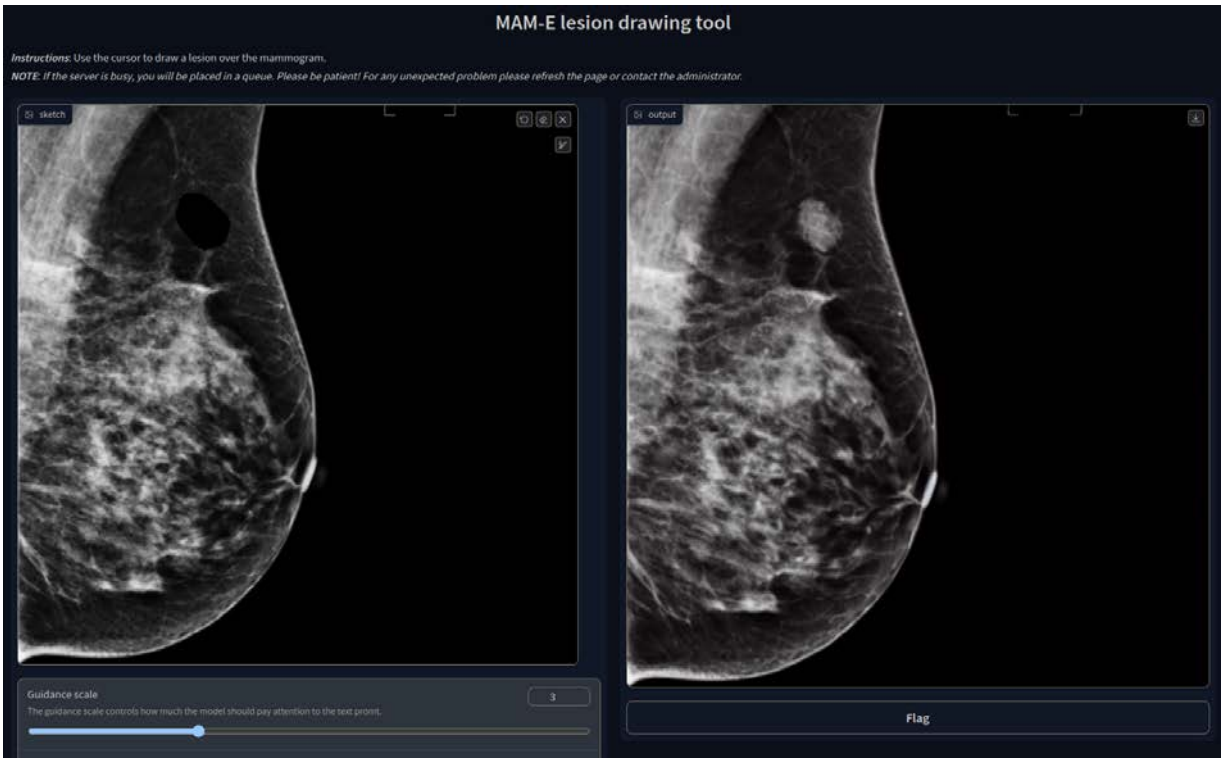


Figure 19: MAM-E for lesion drawing.

GUIs can run in remote servers and be accessible on the web thanks to the *GradIO*, an open-source Python package for rapid generation of visual interface of ML models, by Abid et al. (2019).

We developed five GUIs, one for each of our main diffusion pipelines. Two were designed for the conditional generation of mammograms of the original Siemens and Hologic datasets separately, with their own prompt characteristics. Similarly, one pipeline was created for the fusion of both datasets and it is presented as example in figure 1. In these three cases, the personalization options are set fixed and the user can only pick from the available options. Nevertheless, we added the option of a negative prompt, which allows the user to further personalize the generation.

The idea of the negative prompt is to specify some features that would like to avoid. For instance, in the cases when a synthetic image presents a gray or white background, a negative prompt of "white background" or "no black background" has shown to make the backgorund black.

In the case of the inpainting task, the GUIs has the option to upload the image that will be inpainted, although a default image is available. An interactive drawing brush is then activated, with which a lesion can be inpainted in any part of the mammogram, as shown in figure 19.

Given that the pretrained weights are available in the *Hugigng Face* personal repository of the first author, and that the code to run the GUI interface is publicly available in the GitHub repository of the same authorship, all five GUIs can be run with graphic cards of around 4 GB of GPU memory capacity.

5. Discussion and conclusions

We can encompass the results of this master thesis in three main blocks. The first block consists on exploring the implementation of diffusion models for digital mammography synthesis. The results of the vanilla and pure latent diffusion pipelines show that DM can be adapted for synthetic mammography generation, and that the data distribution of such images can be learned from scratch, although it would require a large dataset and long training time. Indeed, it follows that even though training a DM with one-channel 64x64 images is possible, training a similar model but with four-channels 64x64 latent representations requires more images, GPU resources and time.

Secondly, we found that fine-tuning a SD model pretrained on natural images with mammographic images is feasible and that the objective of the training process reduces to shift the learned data distribution from a non-medical one into the correspondent to our mammography datasets, individually for each mammography unit vendor or combined. This means that we can profit from the essential diffusion properties learned by pretrained natural models which, after trained with huge datasets and for long periods of time, have mastered to denoise images well.

Moreover, we found that stable diffusion text conditioning is a suitable generative model implementation to synthesize mammograms with specific characteristics and properties, giving the possibility to control several aspects of it, such as vendor, view, breast density and breast area. Stable diffusion also opened the possibility of extrapolating characteristics of one dataset into another, thanks to the control given by the CLIP text encoder through attention layers in the UNet, and at inference time by applying classifier-free guidance. We also found that SD can be modified for inpainting of synthetic lesions over healthy mammograms. The developed pipeline essentially only requires the modification of the input latent representation to include a mask to focus the generation process only in that region. All these models inference pipelines were made accessible and ready-to-use through GU interfaces, and the weights and code was made available through personal repositories.

Thirdly, we found initial evidence that the synthetic images coming from our implementation of SD could potentially be used for CAD systems in need of specific image characteristics or with the presence of lesions. A radiological assessment showed that the initial image quality can be compared with real mammograms and the use of explainability AI models helped to explore the behavior of a classification model when tested with our synthetic images with the help of heatmaps.

5.1. Limitations

The first clear limitation of this work is the resolution and pixel depth of the synthetic mammograms. Although at the moment of publication of this work there are some improvements on the SD model for using pretrained weights for 768x768 resolution images, we decided to develop a pipeline first on 512x512 images. This limited resolution reduces the use of our synthetic images on CAD system that require higher resolution, such as micro-calcification detection. The pixel depth was also reduced from its original 16 bits to 8 bits to match the pretrained model requirements. This reduction losses some information in the images and reduces the overall contrast.

Despite the extensive visual assessments perform on the synthetic images, quantitative assessments remained limited in this work. Even though widely used fidelity and diversity metrics, such as the FID score, are being discouraged due apparent model bias, they can still be helpful during training to complement the training performance monitoring. This way, e.g. the model can be stopped or the learning rate can be modified if the generation diversity decreases.

Furthermore, even though some of the SD hyperparameters, such as learning rate and batch size, were changed to explore their effect on the training performance, this work did not prioritized HP tuning and limited itself to HP used in other medical and non-medical DreamBooth implementations.

An important limitation of this work is the lack of deep exploration of the effect of our synthetic images on CAD systems. Even though the assessment of EAI models can give some insights, it is required to train complete CAD pipelines with and without synthetic images to analyze performance changes.

5.2. Future work

Acknowledging the limitations cited above, we plan to explore the use of quantitative metrics during training time, as well as an extensive and organized grid search of the optimal HP for our tasks.

After the release of the pretrained weights for 768x768 resolution images, we expect to perform minimal changes in our current pipeline to allow higher resolution mammography generation.

Additionally, even though the task 1 and task 2 pipelines can be combined manually by directly loading the synthetic or real images into the *MAM-E* drawing tool, we plan to combine this pipeline to fully automatize the generation process in one single GUI.

To assess the training performance of a CAD system using synthetic mammograms, we have started talks with the authors of another ViCOROB lab thesis defendant (Mekonnen et al.) to train a Fast RCNN architecture for lesion detection and bounding box prediction.

Acknowledgments

I would like to thank my master thesis and future PhD supervisor, Robert Martí, for the guidance, patience, flexibility and the opportunity to develop such an interesting research topic in a pleasant work environment. I would also like to thank the ViCOROB community, specially the MAIA fellows, for the mutual camaraderie and the daily motivation to work. A special acknowledgment to the *Hugging Face* organization, for their grandiose labor to make deep learning resources, documentation, courses and source code available for everybody.

Rangiferā tarandā.

Source code

The source code can be found in: https://github.com/Likalto4/diffusion-models_master.
The pretrained weights can be found in: https://huggingface.co/Likalto4.

References

Abid, A., Abdalla, A., Abid, A., Khan, D., Alfozan, A., Zou, J., 2019. Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild. URL: http://arxiv.org/abs/1906.02569. arXiv:1906.02569 [cs, stat].

Chambon, P., Bluethgen, C., Delbrouck, J.B., Van der Sluijs, R., Połacin, M., Chaves, J.M.Z., Abraham, T.M., Purohit, S., Langlotz, C.P., Chaudhari, A., 2022. RoentGen: Vision-Language Foundation Model for Chest X-ray Generation. URL: http://arxiv.org/abs/2211.12737. arXiv:2211.12737 [cs].

Chong, M.J., Forsyth, D., 2020. Effectively Unbiased FID and Inception Score and Where to Find Them, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA. pp. 6069–6078. URL: https://ieeexplore.ieee.org/document/9156949/, doi:10.1109/CVPR42600.2020.00611.

Dhariwal, P., Nichol, A., 2021. Diffusion Models Beat GANs on Image Synthesis. URL: http://arxiv.org/abs/2105.05233. arXiv:2105.05233 [cs, stat].

Dorjsembe, Z., Odonchimed, S., Xiao, F., 2022. Three-Dimensional Medical Image Synthesis with Denoising Diffusion Probabilistic Models.

Frans, K., Soros, L.B., Witkowski, O., 2021. CLIPDraw: Exploring Text-to-Drawing Synthesis through Language-Image Encoders. URL: http://arxiv.org/abs/2106.14843. arXiv:2106.14843 [cs].

Halling-Brown, M.D., Warren, L.M., Ward, D., Lewis, E., Mackenzie, A., Wallis, M.G., Wilkinson, L.S., Given-Wilson, R.M., McAvinchey, R., Young, K.C., 2021. OPTIMAM Mammography Image Database: A Large-Scale Resource of Mammography Images and Clinical Data. Radiology: Artificial Intelligence 3, e200103. URL: http://pubs.rsna.org/doi/10.1148/ryai.2020200103, doi:10.1148/ryai.2020200103.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2018. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. URL: http://arxiv.org/abs/1706.08500. arXiv:1706.08500 [cs, stat].

Ho, J., Jain, A., Abbeel, P., 2020. Denoising Diffusion Probabilistic Models. URL: http://arxiv.org/abs/2006.11239. arXiv:2006.11239 [cs, stat].

Ho, J., Salimans, T., 2022. Classifier-Free Diffusion Guidance. URL: http://arxiv.org/abs/2207.12598. arXiv:2207.12598 [cs].

Kazerouni, A., Aghdam, E.K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., Merhof, D., 2023. Diffusion Models for Medical Image Analysis: A Comprehensive Survey. URL: http://arxiv.org/abs/2211.07804. arXiv:2211.07804 [cs, eess].

Kingma, D.P., Welling, M., 2022. Auto-Encoding Variational Bayes. URL: http://arxiv.org/abs/1312.6114. arXiv:1312.6114 [cs, stat].

Korkinof, D., Rijken, T., O'Neill, M., Yearsley, J., Harvey, H., Glocker, B., 2019. High-Resolution Mammogram Synthesis using Progressive Generative Adversarial Networks. URL: http://arxiv.org/abs/1807.03401. arXiv:1807.03401 [cs].

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J., 2022. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. URL: http://arxiv.org/abs/2206.00927. arXiv:2206.00927 [cs, stat].

Müller-Franzes, G., Niehues, J.M., Khader, F., Arasteh, S.T., Haarburger, C., Kuhl, C., Wang, T., Han, T., Nebelung, S., Kather, J.N., Truhn, D., 2022. Diffusion Probabilistic Models beat GANs on Medical Images. URL: http://arxiv.org/abs/2212.07501. arXiv:2212.07501 [cs, eess].

Nguyen, H.T., Nguyen, H.Q., Pham, H.H., Lam, K., Le, L.T., Dao, M., Vu, V., 2023. VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. URL: http://arxiv.org/abs/2203.11205. arXiv:2203.11205 [cs, eess].

Pinaya, W.H.L., Tudosiu, P.D., Dafflon, J., da Costa, P.F., Fernandez, V., Nachev, P., Ourselin, S., Cardoso, M.J., 2022. Brain Imaging Generation with Latent Diffusion Models. URL: http://arxiv.org/abs/2209.07162. arXiv:2209.07162 [cs, eess, q-bio].

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning Transferable Visual Models From Natural Language Supervision. URL: http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I., 2021. Zero-Shot Text-to-Image Generation. URL: http://arxiv.org/abs/2102.12092. arXiv:2102.12092 [cs].

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-Resolution Image Synthesis with Latent Diffusion Models. URL: http://arxiv.org/abs/2112.10752. arXiv:2112.10752 [cs].

Rouzrokh, P., Khosravi, B., Faghani, S., Moassefi, M., Vahdati, S., Erickson, B.J., 2022. MULTITASK BRAIN TUMOR INPAINTING WITH DIFFUSION MODELS: A METHODOLOGICAL REPORT.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K., 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. URL: http://arxiv.org/abs/2208.12242. arXiv:2208.12242 [cs].

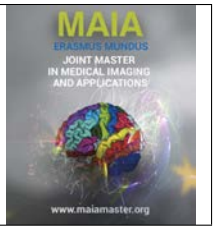
Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S., 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. URL: http://arxiv.org/abs/1503.03585. arXiv:1503.03585 [cond-mat, q-bio, stat].

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention Is All You Need. URL: http://arxiv.org/abs/1706.03762. arXiv:1706.03762 [cs].

Wu, E., Wu, K., Cox, D., Lotter, W., 2018. Conditional Infilling GANs for Data Augmentation in Mammogram Classification, in: Stoyanov, D., Taylor, Z., Kainz, B., Maicas, G., Beichel, R.R., Martel, A., Maier-Hein, L., Bhatia, K., Vercauteren, T., Oktay, O., Carneiro, G., Bradley, A.P., Nascimento, J., Min, H., Brown, M.S., Jacobs, C., Lassen-Schmidt, B., Mori, K., Petersen, J., San José Estépar, R., Schmidt-Richberg, A., Veiga, C. (Eds.), Image Analysis for Moving Organ, Breast, and Thoracic Images. Springer International Publishing, Cham. volume 11040, pp. 98–106. Series Title: Lecture Notes in Computer Science.

Medical Imaging and Applications

Master Thesis, June 2023



Domain specific data augmentation and deep learning architectures for automatic segmentation of the myocardium in delayed enhancement MRI

Gonzalo Esteban Mosquera Rojas, Alain Lalande, Sarah Leclerc

Medical Image Processing team, Institute of Molecular Chemistry of the University of Burgundy (ICMUB), UMR CNRS 6302
University of Burgundy, Dijon, France

Abstract

Delayed Enhancement (DE) cardiovascular MRI is an imaging technique that is acquired some minutes after the injection of a contrast agent. Automatic segmentation on such images is a topic that has raised a lot of attention in the medical imaging community, since it is strongly involved in myocardium viability assessment, which refers to the amount of tissue that recovered its functionality after undergoing revascularization therapy due to a previous event of myocardial infarction (MI). It has been previously found that automatic segmentation models struggle a lot to segment the myocardium when they face cases of MI, since these areas usually showcase an irregular and heterogeneous aspect in terms of shape and intensity, and can also partially obstruct the view of this structure. To overcome this issue, we propose an image processing based data augmentation algorithm where we create synthetic cases of myocardial infarction from healthy ones. The algorithm relies on prior information extracted from an external dataset where MI cases are available, and is able to automatically generate new MI samples with varying size, type and location. The method can be applied under two different scenarios: a fixed generation and an adaptive one. In the first scenario, the training dataset is enlarged with any previously defined rate whereas in the second the algorithm collects feedback during model training and perform the data augmentation exclusively on difficult cases. We evaluate the impact of approaching the problem under two data scenarios, single modality and multi modality. In this latter, information from kinetic (CINE) MRI, which is an image that is also acquired along DE in a typical cardiovascular examination, is also exploited by the model, and the extracted features are fused at an intermediate step. The results show that addressing the problem in a multi modal fashion and adding the data augmentation algorithm leads to a more consistent segmentation of the myocardium in DE MRI, as the model is able to relate the MI areas with the myocardium, thus increasing its overall robustness to pathology specific local pattern perturbations.

Keywords: Cardiac imaging, myocardial infarction, delayed enhancement MRI, image segmentation, data augmentation

1. Introduction

According to the World's Health Organization (WHO), cardiovascular diseases (CVDs) are one of the main causes of global mortality. In 2005, they accounted for 17 out of 58 million worldwide deaths, from which 7.6 million were due to coronary heart disease (CHD) (Mendis et al., 2011).

Myocardial infarction (MI), also known as heart attack, is one of the ways in which CHD can appear (Mendis et al., 2011). This condition refers to the damage or death of a certain portion of the heart muscle (myocardium) due to an unexpected interruption or decrease of blood supply (ischemia), which is caused by blocks in the coronary arteries. Although they can vary widely among patients, the symptoms of MI typically include chest pain that can spread up to other parts of the body, as well as a general sensation of fatigue and discomfort. Myocardial infarction can be divided in five different groups depending on pathological, clinical and prognostic factors (Thygesen et al., 2012).

MI of type 1, also known as spontaneous MI, is associated with atherosclerotic plaque rupture, generating thrombus in one or several coronary arteries. This con-

dition can be caused by the presence of Coronary Artery Disease (CAD), but not necessarily. Type 2, or MI secondary to an ischemic imbalance, as it name suggests, occurs due to a disparity between oxygen supply and demand caused from underlying pathological conditions different to CAD, e.g, hypertension, respiratory failure, anemia, etc. MI of type 3 is considered responsible of cardiac death. Typically, patients would present symptoms associated with myocardial ischemia, but they die before any blood tests or analyses can be performed. MI of type 4 is subdivided in two categories: 4a and 4b. The first one is related to percutaneous coronary intervention (PCI), while the second is liked to stent thrombosis. The 5th type of MI is related to coronary artery bypass grafting (CABG) (Thygesen et al., 2012).

Nowadays, there are several strategies used in medical practice for MI diagnosis, namely, analysis of electrocardiogram (ECG), tracking of the levels of biomarkers that could suggest potential myocardial tissue damage or death, and medical imaging. This latter embodies different techniques which include echocardiography, scintigraphy, magnetic resonance imaging (MRI), and computer tomography (CT) (Thygesen et al., 2012).

After the heart has suffered from an event of MI, it is essential to evaluate the viability of the myocardium (Lalande et al., 2020). According to Thomson et al. (2004), one of the defacto definitions of viability is "recovery of contractile function following revascularization", being revascularization the name of a group of procedures whose goal is to restore blood flow from blocked arteries. In medical practice, there are two main cardiac MRI techniques for evaluating the structure and function of the heart, namely, kinetic (CINE) and DE (Delayed Enhancement) MRI. The latter, as its name suggests, is performed some minutes after the injection of a contrast agent (Lalande et al., 2020). Due to its nature, CINE MRI is more suitable for performing heart motion and contraction analysis both locally and globally, whereas DE MRI provides useful insights to assess myocardial tissue damage, since the contrast agent allows these areas to be highlighted. These images are typically acquired in a short axis view, which is one of the standard imaging planes for cardiac MRI. In this type of view, the imaging plane is positioned perpendicular to the long axis of the heart, slicing through the ventricles at various levels from base to apex, thus allowing a thorough assessment of the heart structure.

The analysis of the MI extent implicitly requires the segmentation of the myocardium in DE MRI, which then involves the segmentation of the left ventricle cavity. This is a topic that has raised a lot of attention in the Medical Image Analysis community, since it can be very challenging due to different factors such as artifacts and noise in the image, but most importantly due to the varying contrast between normal and abnormal myocardial tissue, since the MI appearance ranges from being bright and well defined to being subtle and heterogeneous in some other cases. International challenges such as EMIDEC in 2020 validate the current need of having reliable and robust automatic methods to perform this task.

When assessing myocardial viability, both MRI modalities can offer complementary information, resulting in a thorough evaluation of the state of the heart that takes into account both its anatomical and functional aspects. In the work of Ouadah et al. (2022) they found it useful to cross information from both modalities for the segmentation of the myocardium in DE-MRI by using an intermediate fusion scheme, and also determined that the model struggled to segment cases in which myocardial infarction was present.

The main objective of this work was to build a deep learning based robust pipeline that could improve the overall quality of the segmentation of the myocardium in DE-MRI in order to make useful for clinical application. The problem was addressed under the following scheme: first, we propose and implement a data augmentation algorithm in which synthetic myocardial infarction samples are created on healthy myocardium with the goal of providing the segmentation models with diverse pathological data so that they can further learn from patterns in these cases. Secondly, the algorithm is applied under two scenarios: fixed and adaptive. In the first one, the training set is enlarged at different rates and then fed to the segmentation model. On the other hand, the adaptive data augmentation, as it name suggests, dynamically gets feedback from the model during training to identify the cases where it is struggling the most and new synthetic data samples are created from these particular cases.

The work is validated on single modality and multi modality settings, using a modified version of the traditional UNet architecture proposed by Ronneberger et al. (2015), and an intermediate-fusion UNet architecture investigated in the work of Ouadah et al. (2022), respectively. These two architectures were defined as the baseline models for the whole study. In the final part of the work, some additional experiments were done using the nnUNet framework, proposed by Isensee et al. (2021), to perform the segmentation task. This was done with the goal of evaluating the performance of the framework itself, which has proven to be quite robust for medical image segmentation, but also to extract some useful insights to define future directions of work to further increase the robustness of the pipeline for this particular application.

2. State of the art

Semantic segmentation is typically defined as the process of dividing an image in a group of regions that correspond to objects of interest. In the cardiac imaging field, these regions are associated to important anatomical structures such as the ventricles, myocardium, coro-

nary arteries, etc. In order to differentiate such objects, the segmentation process involves performing a thorough analysis of the image in terms of shape, pixel intensity and texture.

Image segmentation plays a key role in cardiology since it allows doctors to do a much more in-detail assessment of the heart function, e.g, measuring ventricular volume, ejection fraction, myocardial viability, wall motion, etc. Therefore, it has been a topic of great interest for researchers in the medical imaging domain, with the goal of providing computer-based solutions that can be robust and reliable for clinicians to use in daily practice.

The following subsections aim to present the literature revised in the development of this work in order to understand the current state of the art of deep learning based methods for the segmentation task within the field of cardiac imaging.

2.1. CINE MRI segmentation

In the work of Petitjean and Dacher (2011), a comprehensive review of segmentation methods in short axis cardiac MRI was presented. The revised models were classified in two main categories, namely, methods based on none or weak priors, and methods based on strong priors. The first group consisted on image and pixel classification based methods, as well as deformation models. On the other hand, the second category included shape prior based deformable models, active shape and appearance models, and atlas based methods. They conclude that, in general, the results for left ventricle segmentation are satisfactory for mid ventricular slices, and the improvement is rather constrained to focusing on the basal and apical ones.

Nonetheless, the problem of automatic segmentation of cardiac structures in CINE MRI has also been a research topic in the deep learning field. One of the most well-known challenges in this regard is the Automated Cardiac Diagnosis Challenge (ACDC) proposed by Bernard et al. (2018) within the framework of the MICCAI conference 2017 edition. The dataset is composed of sequences acquired from 150 patients, obtained over 6 years with MRI scanners with 1.5T and 3.0T of magnetic strength. The dataset is evenly distributed in 5 classes, namely, patients with normal cardiac anatomy and function (NOR), patients with a systolic heart failure with infarction (MINF), patients with dilated cardiomyopathy (DCM), patients with hypertrophic cardiomiopathy (HCM) and patients with abnormal right ventricle (ARV). Therefore, the competition also aimed to address the problem of classification of the examinations. From the total amount of samples, 100 and 50 are left for training and testing purposes, respectively (Bernard et al., 2018).

From the results of the challenge, the work of Isensee et al. (2018) consistently obtained the first place. For the segmentation task, their strategy consisted on using an ensemble of 2D and 3D models, which correspond to modified versions of the traditional UNet architecture, proposed by Ronneberger et al. (2015). They addressed the problem of varying data resolution by resampling the volumes according to each model as a preprocessing step (Isensee et al., 2018). The 3D model was composed by two paths, one for context aggregation and another one for localization. Both are linked at different scales with the goal of merging contextual features with local ones. Models in both dimensions are completely equivalent and just change the dimension-dependent operations accordingly (Isensee et al., 2018). They obtained a mean Dice score of 0.967, 0.946, and 0.896 for the left ventricular cavity, right ventricular cavity and myocardium, respectively, in the end diastolic phase, and a score of 0.928, 0.904 and 0.919 for the same respective structures in the end systolic phase.

In the same competition, the work of Zotti et al. (2018) ranked second. The main strength of their pipeline lies on the segmentation architecture itself. The model, which they call "GridNet" is an extension of UNet. It consists of three columns and five rows. In each row, features are extracted at different scales. The scale from row n corresponds to a down sampled version of scale n-1 by a factor of two. In the case of the columns, these contain convolutional operations to compute features at several resolution levels. However, the last column is used for aggregating the previously computed features in ascending resolution. The output from the last column is then concatenated with a shape prior model, which they generated by estimating the probability of a 3D location to belong to any of the classes of interest (Zotti et al., 2018). They obtained a mean Dice score of 0.964, 0.934, and 0.886 for the left ventricular cavity, right ventricular cavity and myocardium, respectively, in the end diastolic phase, and a score of 0.912, 0.885 and 0.919 for the same respective structures in the end systolic phase.

In general, most of the teams that took part of the challenge based their pipelines on the UNet architecture in its 2D and 3D versions and proposed their own modifications either in the pre-processing, postprocessing or the architecture itself. However, a few teams used some other approaches to tackle the problem. For instance, in the work of Rohé et al. (2018) they propose the use of a Stationary Velocity Field newtork (SVFNet), which learns and predicts the deformation between pairs of images. Their approach consists on aligning all images to a common position, then registering template images to their corresponding target through the SVF Net, and finally propagating the labels and fuse them to get the final prediction. The method of Wolterink et al. (2018) is based on a CNN that uses dilated convolutions. Finally, the work of Tziritas and Grinias (2017) tackled the segmentation problem with an opimized version of Markov Random Fields (MRF).

2.2. DE MRI segmentation addressed as a single modality approach

Although CINE and DE MRI share some intrinsic characteristics, segmentation on this latter is quite challenging due to the fact that it might have lower spatial resolution and scar regions (which are within the myocardium) usually showcase an irregular and heterogeneous aspect in terms of shape, intensity and size. These limitations have made DE MRI segmentation a process that requires manual intervention by trained experts, which is both time consuming and subject to inter and intra observer variability. Therefore, this topic has raised a lot of attention in the deep learning community, with the main goal of building automatic methods that can perform this task in a robust and accurate fashion.

One of the most popular challenges for DE-MRI segmentation is EMIDEC, proposed by Lalande et al. (2020) within the MICCAI conference in its 2020 edition. The dataset is composed of 150 DE-MRI examinations in short axis orientation with its corresponding ground truth and a text file with clinical information from each patient. From the total amount of available exams, one third of them correspond to normal cases and two thirds to patients with acute MI. The challenge involved both segmentation and classification tasks. For the first one, the main goal was to delineate the myocardium, infarction tissue and permanent microvascular obstruction area (PMO).

From the results of the challenge, which were reported by Lalande et al. (2022), the work of Zhang (2021) consistently ranked in the first place for the segmentation task in all the structures of interest. Their pipeline consists on a cascaded CNN which is divided in two parts. In the first one, they unroll the volume to their corresponding slices and feed it to a 2D UNet in order to obtain what they call a "coarse segmentation", which has been learned from intra slice information (Zhang, 2021). In the second stage, the original input is concatenated with the coarse segmentation results and they are then fed to a 3D UNet. With this strategy, they make sure that the segmentation results are more robust and accurate since the whole pipeline is also being able to learn inter-slice variability and spatial context from the original volumes. They obtained a Dice score of 0.879, 0.712, and 0.785 for the myocardium, infarction area and PMO, respectively. The implementation of this method was based on the nnUNet framework, proposed by Isensee et al. (2021). In summary, this is a deep learning segmentation method that is able to configure itself in all the stages that involve solving this task, namely, preprocessing, network architecture, training procedure and postprocessing (Isensee et al., 2021). In terms of the segmentation architecture, it offers the 2D and 3D versions of the UNet. For the latter, a low and full resolution are available depending on dataset findings, and also a cascaded version in which a first 3D network learns from downsampled data, and then the predictions are upsambled back to the original resolution and concatenated (in one-hot encoding form) with the original data (Isensee et al., 2021).

Within the framework of EMIDEC challenge, the work of Yang and Wang (2021) was also revised. It consists on a hybrid version of the UNet architecture in which they replace the encoder and the decoder parts for an squeeze and excitation residual (SE-REs) module and a selective kernel (SK) block, respectively. The first module allows the model to learn from existing relationships between feature channels and focus the training on those who are the most informative, whereas the second block dynamically changes the receptive field size to capture feature information at multiple scales. Since their segmentation network is 2D, they extract inter slice features by creating an image where the three neighboring slices from a given slice are stacked and fed into the model. They obtained a Dice score of 0.855, 0.628 and 0.610 for the myocardium, infarction area and PMO, respectively.

2.3. DE-MRI segmentation addressed as a multi-modality approach

In a typical MRI exam, both CINE and DE protocols are acquired. Therefore, when doctors need to analyze DE MRI, they usually extract some supporting information from its paired CINE image, where anatomical structures, such as the myocardium, are more visible(Dikici et al., 2004). It is thus natural to think that, in accordance to medical practice, leveraging from the information of CINE MRI might have a positive impact when building automatic segmentation pipelines on DE MRI.

In the work of Dikici et al. (2004), whose goal was to perform a quantification of non-viable tissue in DE MRI, they perform an automatic segmentation on the corresponding CINE MRI and use it as a segmentation prior that is iteratively deformed to maximize the overlap with the correct segmentation on DE MRI. This fitting is done through an affine registration procedure that includes translation, shearing and scaling operations. Finally, they perform the myocardial pixels classification (viable vs non-viable) using an SVM classifier.

The method proposed by Ciofolo et al. (2008) consists on three main stages. In the first one, a geometrical template is initialized for each of the DE MRI slices and then deformed to match a previously defined model for the myocardium contour. Then, a 3D mesh is built from the myocardium contours of CINE MRI. They propose the use of a 3D mesh since it allows to represent both the myocardium geometry and thickness. In the final step, the mesh is registered to the 2D space of the contours found in the first stage.

Even though are image processing based solutions, there are also a few that used deep learning. For instance, in the EMIDEC challenge, the segmentation

pipeline of Huellebrand et al. (2021) proposed three different CNN-based models. One of them was merely trained on the challenge data, the other one was previously trained on an external dataset of 100 DE-MRI examinations and the third, which achieved the best results, is based on a hybrid mixture model and uses both the additional and challenge data for training. In all the experiments they conducted, the common factor was the use of transfer learning, as they used as baseline a pretrained CNN for the segmentation of CINE-MRI on the ACDC challenge, and it proved to have a positive impact on the final segmentation results.

In the work of Ouadah et al. (2022), the task of DE MRI segmentation was addressed by the concept of data fusion between CINE and DE MRI. Their goal was to evaluate whether adding information from CINE was beneficial for the segmentation on DE, and if so, which data fusion approach performed best. The analyzed strategies included input, output, layers, intermediate and self supervised module adaptation fusion. Their results showed that leveraging from a clearer spatial and boundaries information in CINE had a positive effect on the structures segmentation on DE, and also the UNet with an intermediate fusion scheme (which they called "DualUNet") showed to perform best. This fusion scheme was inspired by the work of Xue et al. (2020), where T1 and T1 flipped brain scan images were fused, before the decoding path, for stroke lesion segmentation.

2.4. Data augmentation strategies in cardiac MRI segmentation

Although deep learning based segmentation methods have shown promising results in a variety of different applications, they do need big amounts of data in order to generalize well. This is quite problematic in the field of medical imaging, since the amount of available data is generally limited, and the annotation task is complex and costly.

Data augmentation emerged as a technique to solve the above mentioned problem. As it name suggests, it involves generating new data samples by applying several transformations on the existing one. Some of the most common transformations reported in the literature include rotations, translations, scaling and deformations. Additionally, there is a group of deep learning methods, commonly known as generative models, which can automatically create synthetic images, and thus have also been used as data augmentation tools.

There have been some previous works where data augmentation has been used for the task of segmentation in cardiac MRI. For instance, Chen et al. (2022) proposed an adversarial data augmentation pipeline. Generally speaking, this concept refers to the process of applying a set of different perturbations on the original images to try to fool the model. In a posterior step, these augmented data samples are used during training to allow the model to learn from a more varied version of the data and make it robust to perturbations. Their method can dynamically optimize all the transformation parameters and thus produce result that are realistic and plausible in typical medical imaging setups.

Lin et al. (2020) proposed a shape-based data augmentation algorithm where they learn different representations of the left ventricle and capture structural relationships between shapes. In summary, given a dataset of images with their corresponding left ventricle contours, one image is selected as reference and the contours of the rest are deformed to match the reference one. The transformation parameters are stored and taken as shape features. In posterior steps, new shapes are generated from an orthonormal feature basis that is computed based on eigenvector analysis on the correlation matrix of shape features (Lin et al., 2020).

Finally, Skandarani et al. (2020) proposed the use of Variational Autoencoders and Generative Adversarial Networks (GANs), originally proposed by Goodfellow et al. (2020), for generating realistic synthetic MRI. Their results show that training Convolutional Neural Networks with the data generated by their model achieved competitive performance with respect to other traditional techniques.

3. Material and methods

3.1. Datasets

3.1.1. CINEDE dataset

CINEDE was the main dataset source of this work, i.e., where the segmentation models were built upon. It consists of CINE and DE MRI examinations of 124 patients, as shown in figure 1. The exams were acquired in the University Hospital of Dijon, France, using 1.5T and 3T magnets (Siemens Medical Solution, Erlangen, Germany), with a phased toracic coil. DE-MRI was acquired 10 minutes after the injection of a gadolinium-based contrast agent at a concentration of

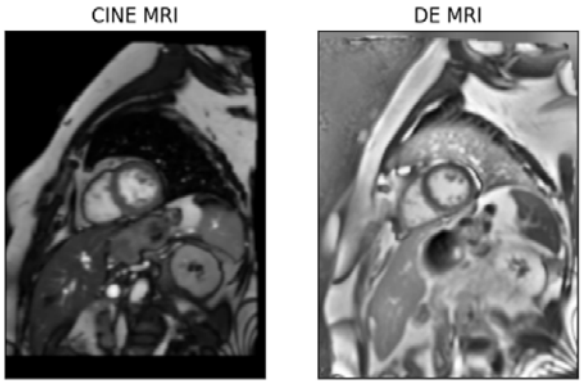


Figure 1: CINE and DE MRI slices corresponding to the same patient.

0.1 or 0.2 mmol/kg. The final volumes are generated in NIFTI format, and represent a short axis view of the left ventricle, covering basal, middle and apical slices, and with a number of slices per volume that varies between 7 and 12, with a total of 984 2D slices. The median voxel spacing in x,y, and z axes for the CINE and DE modalities are $[1.36719, 1.36719, 10]mm^3$ and $[1.875, 1.875, 10]mm^3$, respectively. All images have their corresponding manual annotations (ground truth) with two labels: left ventricle cavity and myocardium, and were acquired with a short axis view of the heart.

The dataset in itself is quite challenging since it represents cases from five different contitions:

- Dilated cardiomyopathy (CMD), associated with an enlargement and weakening of the ventricles.
- Normal MRI exams (NOR)
- Myocarditis (MYO), which refers to the inflamation of the myocardium.
- Other pathologies (OTH), associated with rare diseases with low presence in the dataset.
- Myocardial infarction (VIA), referring to the death of a certain portion of the myocardium tissue.
- Hyperthrophic cardiomiopathy (CMH), associated with an abnomal high thickness of the myocardium.

This diversity in the dataset translates to large set of distinct features that the model must learn as they might have impact on how the anatomical structures appear, thus increasing the difficulty of the segmentation task. Moreover, as figure 2 depicts, the samples distribution among different classes is uneven.

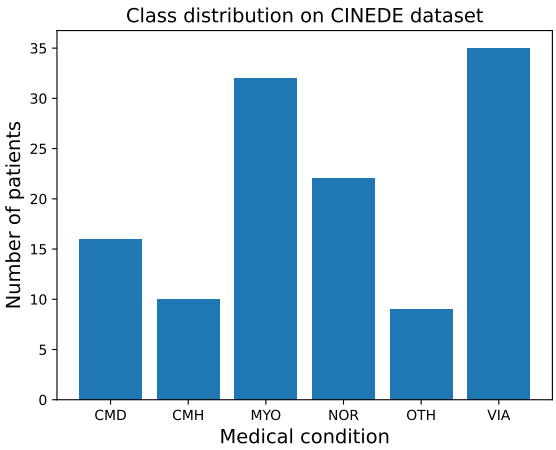


Figure 2: CINEDE dataset class distribution.

3.1.2. EMIDEC dataset

EMIDEC is an auxiliary dataset that was used for the extraction of prior information that is fed into our data augmentation algorithm, proposed by Lalande et al. (2020) within the framework of the automatic evaluation of myocardial infarction from DE MRI challenge. It consists of 150 DE MRI examinations with a text file that contains clinical information from the patient, and the corresponding ground truth image that maps four labels: left ventricle cavity, myocardium, myocardial infarction and persistent microvascular obstruction (MVO) area. The acquistion protocols were similar to those from CINEDE. The data samples distribution between normal and pathological cases is one third and two thirds, respectively. All abnormal cases correspond to patients with acute myocardial infarction. The pixel spacing of the examinations ranges from 1.25 x 1.25 mm^2 to 2x2 mm^2, the slices have a thickness of 8 mm and the distance between slices is 10 mm(Lalande et al., 2020).

3.2. Method

Figure 3 presents a summary of the steps that were performed to approach the target problem, which are described in the following sections.

3.2.1. Data preprocessing

Generally speaking, data preprocessing plays a crucial role when training deep learning models, since they require data to be as clean as possible for an effective learning process. For this particular problem, the preprocessed consisted of several stages. First, images were normalized within the rage of 0-255. Afterwards, the labels of raw images were converted to integer values, since they were initially reported as floating numbers. In third place, the orientation of all images was fixed to left posterior inferior (LPI). As a fourth step, the size of both images in a given pair were made equal by zero-padding the smaller one. Afterwards, the spatial correspondence between the two modalities was uniformed with the goal of having consistent contextual information across modalities. This was achieved by resizing all images to match the median voxel spacing of DE MRI, as it is the target modality in this study. As a final step, CINE MRI was registered to DE MRI to make them suitable for fusion based deep learning architectures under the multi modality setting.

3.2.2. Segmentation architectures

The problem was addressed under two different scenarios: single and multi modality. As their name suggest, in the first one only the information from the target modality (DE-MRI) was taken into account, while in the second one CINE-MRI was also considered, so the

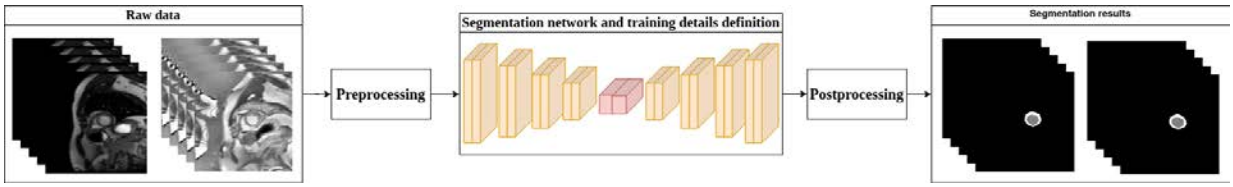


Figure 3: General workflow summary.

data becomes of paired nature. For the single modality approach, the UNet architecture, proposed by Ronneberger et al. (2015), was used. Batch normalization layers were added in the building blocks of the network.

For the multi-modality approach, the "DualUNet" architecture, which was investigated in the work of Ouadah et al. (2022), was used. This architecture consists on a modified version of the original UNet but it allows to do data fusion between the CINE and DE MRI at an intermediate step. As the figure 4 shows, one independent UNet like encoder path is used for each modality in order to extract separate distinct features. The output of the convolution blocks at level i are combined through a fusion block in which several operations take place, as depicted in figure 5. First, both feature maps are stacked, then fusion features are calculated through a 3D convolutional layer, and the result is taken back to the original feature map dimension via squeeze operation.

For the final additional experiments, the segmentation task was also tackled using a customized version of the nnUNet framework from Isensee et al. (2021) in both single modality and multi modality settings.

3.2.3. Extraction of prior information from EMIDEC dataset

Ouadah et al. (2022) found that segmentation models tend to have the worst performance in cases where myocardial infarction is present. Hence, in this work we propose a data augmentation algorithm in which synthetic infarction cases are generated from images of healthy patients. Afterwards, these newly created images are added to the training set and are fed to the segmentation models, with the goal of guiding them towards a better learning of this anatomical structure.

The algorithm feeds itself from some prior information in terms of type, location, area and intensity in order to generate cases that are as much realistic as possible. These priors are extracted from the pathological cases of EMIDEC dataset, which, as mentioned in section 3.1.2, completely correspond to MI cases. The steps for evaluating the priors were the following:

Type: In an algorithmic point of view, the myocardial infarction was constrained to have two types: it either crosses the whole myocardium (transmural MI, called TMI hereinafter) or not (non-transmural MI, called NTMI hereinafter). Figure 6 shows an example of the masks for these two types of MI.

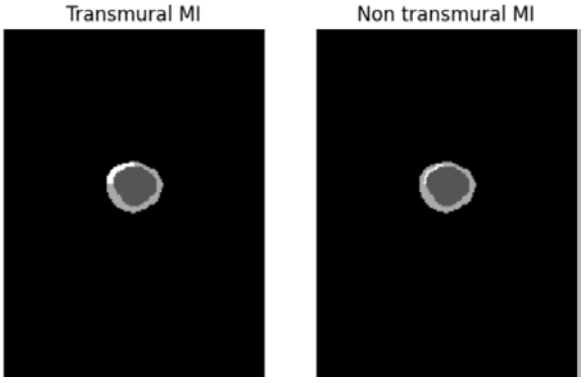


Figure 6: Examples of transmural and non transmural MI. Dark gray: left ventricle cavity, light gray: myocardium, white: myocardial infarction.

Location: To understand the typical location of MI on real world data cases, all images from EMIDEC dataset were divided in four equal parts or "quadrants", being quadrant 1,2,3 and 4 the ones in the top left, top right, bottom left and bottom right location, respectively. In order to do a thorough analysis in this regard, a modified version of the "Bull's Eye Plot" was constructed for the EMIDEC dataset. Generally, this is a figure composed by four concentric circles that are divided so as to represent the 17 segments of the heart proposed by Cerqueira et al. (2002), and it is used to study heart functioning features such as myocardial perfusion and wall motion. Each portion of the graph is colored following a heat map that is generated with all the per segment values of the cardiac feature that is being analyzed.

The modified version of the Bull's Eye plot is presented in figure 7. The yellow ring is a control structure that represents the total amount of the population in the dataset. The rest of rings represent, from the innermost to the outermost, the basal, mid and apical slices of the examination volumes, respectively. These are divided in four parts, which correspond to the four quadrants that were defined for the extraction of location priors. This sums up to a total of 13 parts, as opposed to the 17 present in the traditional plot. The main goal with modifying the graph was to do a more generic analysis

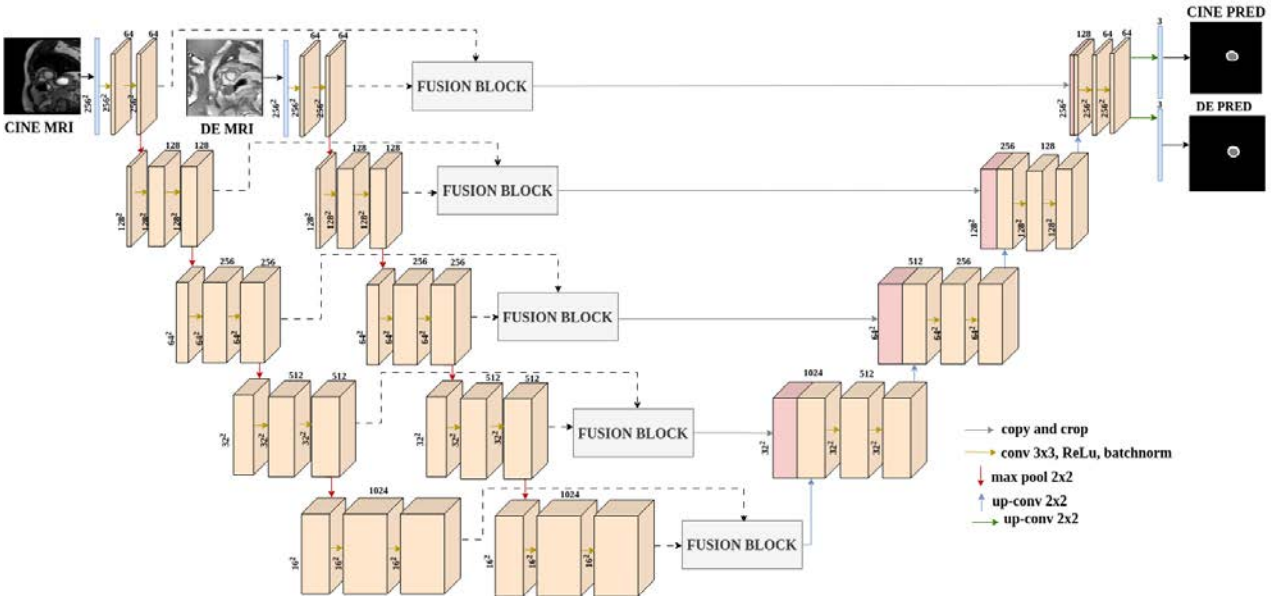


Figure 4: DualUNet architecture.

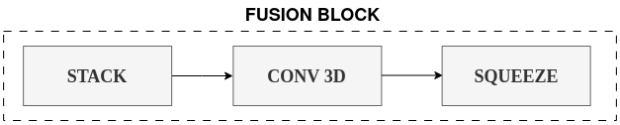


Figure 5: Fusion block of DualUNet architecture.

at the slice level rather than at the segment level. The color of each portion of the plot represents the amount of patients within the EMIDEC dataset for which most of the myocardial infarction was located at a given slice view and quadrant. From the graph it can be seen that, although there is a higher prevalence of the MI in the first quadrant and mid and basal slices, the color variation is rather subtle, which means that the difference is not big enough to robustly establish a location pattern of the MI structure. Hence, based on this evidence, it was decided to keep the location prior of the algorithm as a random variable that can take values between 1 and 4 with equal probability.

Size: The size prior was defined in terms of which percentage of the total amount of myocardium corresponded to infarction tissue. The distribution of this quantity over all patients is depicted in figure 8. It can be seen that, for 50% of the patients, which is the inter quartile range of the box plot, the MI tissue occupied between 12% to 29% of the myocardium. However, for setting the priors in the data augmentation algorithm, most of the possible range of values were taken into account. As a result, the size prior for the MI was defined as a portion of the total myocardium that ranges from 10% to 50%, with equal probability of selection.

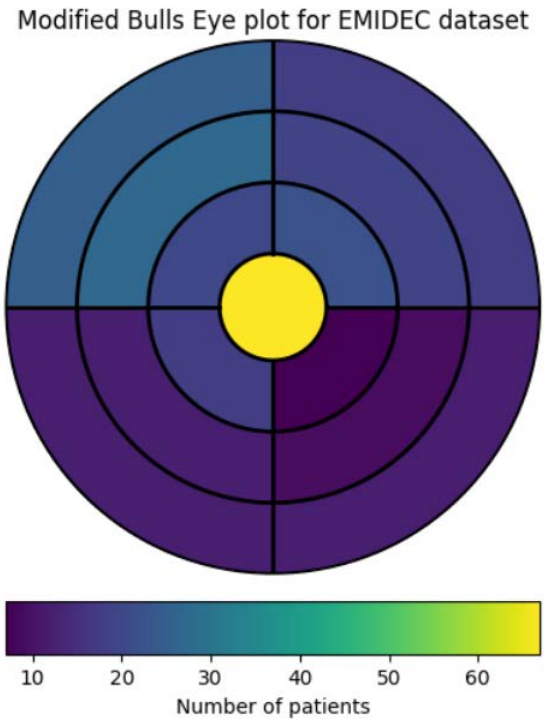


Figure 7: Modified Bull's Eye plot for EMIDEC dataset.

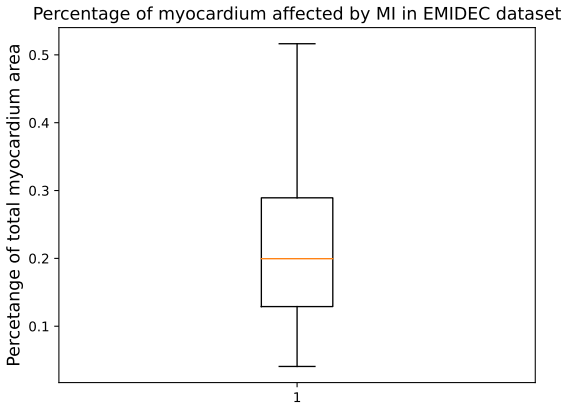


Figure 8: Distribution of the percentage myocardium area affected by MI in EMIDEC dataset.

Intensity: Once the myocardium mask is generated using the above mentioned priors, the corresponding perturbation must be done on the MRI to generate the synthetic image that is to be used to enlarge the training set and provide the models with more samples of MI. For each of the MRI examinations on CINEDE, the intensity of the MI area was defined as the mean intensity of the left ventricle cavity plus a prior knowledge value. This value is sampled from a generated normal distribution whose mean is the average of the differences between the maximum intensity value of the MI (red contour on figure 9) minus the mean value of the left ventricle cavity (blue contour on figure 9) across the entire pathological cases on EMIDEC dataset, with the goal of extracting contrast information between those two anatomical structures. The selection of a normal distribution is done to ensure a consistent MI shape in terms of intensity, and the starting value is selected as the mean intensity of the left ventricle since it was also found that the MI is always brighter.

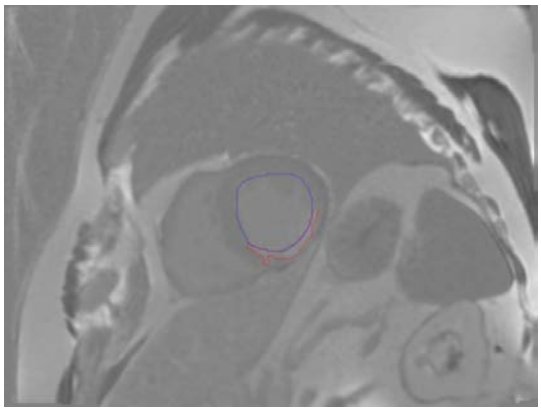


Figure 9: Example of pathological case on EMIDEC dataset (Lalande et al., 2020). The red and blue contours correspond to the MI and left ventricle cavity, respectively.

3.2.4. Data augmentation approach

Once the priors are defined, the data augmentation algorithm consists of the following steps:

Initialization: Given the original image and its corresponding ground truth as inputs, the type and location of MI are defined as a random number picked from the range $\{1, 2\}$ and $\{1, 2, 3, 4\}$, respectively. The size of the MI corresponds to a percentage sampled randomly from a uniform distribution within the range [0.1,0.5]. The algorithm then calculates the effective MI size as a multiplication of the total myocardium pixels by the percentage that was randomly sampled.

Extraction of contours: As a second step, the contours for the left ventricle cavity and the myocardium are extracted, as shown on image b from figure 10.

Division of contours in quadrants: Once the contours have been extracted successfully, four binary images are generated, which contain the parts of both contours that fall within the range of coordinates of the previously defined quadrants. As a final step, the binary image where the MI is to be generated is kept, as depicted in image c from figure 10

Definition of boundary points: As a fourth step, one random pair of coordinates from the left ventricle cavity contour, located within the selected quadrant, is chosen as the starting point (*start CV*) for generating the MI shape. Afterwards, the algorithm selects the left ventricle cavity end point (*end CV*) as the one whose distance with *start CV* is the closest to the effective MI size. This is a design simplification that was set since the size prior is quite flexible and has a big range of possible values, thus it was not a constraint to exactly match the total expected size, as the main objective is to generate different shapes of MI for the segmentation models to learn from them. The starting and end points of the myocardium (*start myo* and *end myo*, respectively) correspond to the closest points (measured in euclidean distance) to *start CV* and *end CV*, respectively. The result from this step is presented in image d from figure 10, where the white arrows indicate the four boundary points.

MI contour generation (TMI): As a first step, all the coordinate pairs falling between the start and ending points of each structure (myocardium and left ventricle cavity) are given the label 3 (corresponding to MI), and the results are shown in image e from figure 10. In order to close the contour, one line is traced between *start LV* and *start myo* and another one between *end LV* and *end myo*, as demonstrated by image f in figure 10. This is done through Bresenham's line algorithm, proposed by Bresenham (1965), which is widely used in computer graphics to draw lines between two points in a discrete grid. In summary, the main idea of the algorithm is to make decisions on whether to step horizontally or vertically at each pixel in order to get the closest approximation to the ideal line. In algorithm 1 a pseudo code of this method is presented for the case in which the line

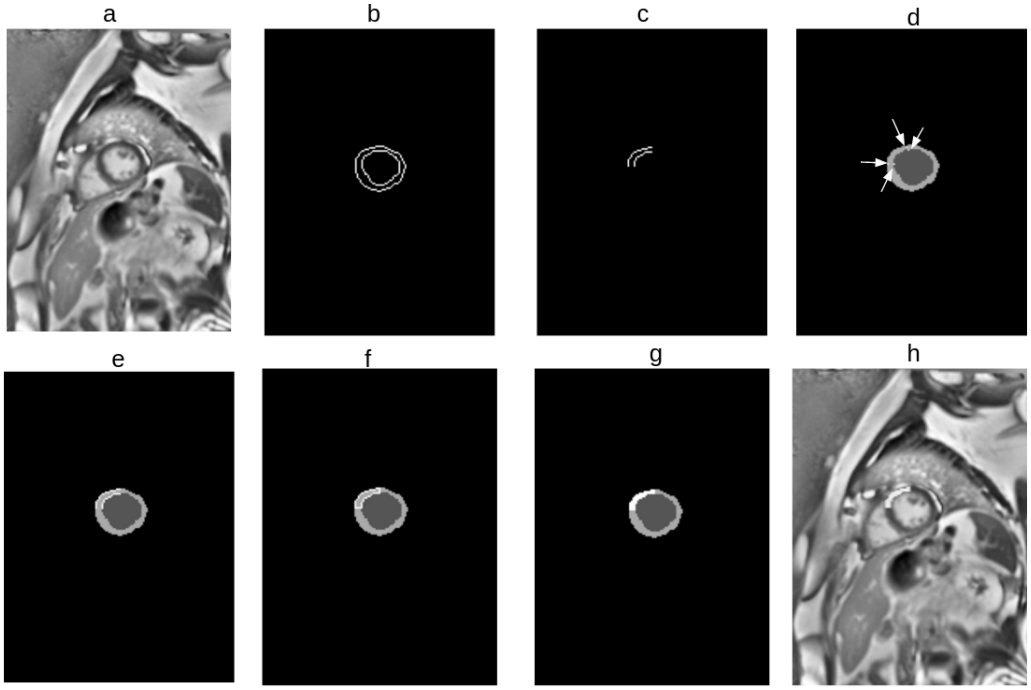


Figure 10: Step-by-step summary of the data augmentation algorithm for a case of MI type 1, a: original DE MRI, b: detection of contours, c: selection of working quadrant, d: definition of boundary points, e: coloring of points within the boundaries range, f: contour closing with Bresenham's algorithm, g: contour filling, h: generation of MI in DE MRI.

goes from left to right and has a slope between 0 and 1. It can be observed the decision variable D gets updated in every iteration of the loop and its value allows the algorithm to decide whether to increment the y variable by one or leave it unchanged. At the end of each iteration, the image is assigned the label of the MI to close the contour. All the other line cases can be covered with subtle modifications of this pseudo code.

MI contour generation (NTMI): Once the four boundary points have been calculated, only *start CV* and *end CV* are kept, and the points belonging to the CV contour and falling between them are assigned label 3. Since this type of MI does not cross the entire myocardium, the second pair of points are no longer part of the myocardium contour. However, *start myo* and *end myo* are used as a guide, and all the points of the line that would be traced between them if it was a TMI are stored in an array called *candidate points*.

For this type of MI the variable thickness is created. This is defined as a percentage of the total myocardium width, and is automatically sampled from a random distribution that ranges between [0.2,0.8]. This quantity is then multiplied by the myocardium width to get the effective desired width of the MI to be created. Hence, the end coordinate pairs in this case are those from *candidate points* whose distance is the closest to the desired width.

Similarly to TMI, a line is traced between starting and end points of both sides of the shape. However, since the

Algorithm 1 Bresenham's line algorithm applied to close MI contour

function BRESENHAM($image, x_0, y_0, x_1, y_1$)
being (x_0, y_0) and (x_1, y_1) (start CV, end CV) or (start myo, end myo)
$\Delta x \leftarrow x_1 - x_0$
$\Delta y \leftarrow y_1 - y_0$
$D \leftarrow 2\Delta y - \Delta x$
$x \leftarrow x_0$
$y \leftarrow y_0$
$image(x, y) \leftarrow 3$ ▷ Assigns MI label to close contour
while $x < x1$ **do**
if $D < 0$ **then**
$D \leftarrow D + 2\Delta y$
else
$y \leftarrow y + 1$
$D \leftarrow D + 2(\Delta y - \Delta x)$
end if
$image(x, y) \leftarrow 3$ ▷ Assigns MI label to close contour
end while
end function

end points are no longer part of the myocardium contour in this case, a connection between them must be generated. Due to the randomness of the algorithm, it cannot be assured that the connection using a line will be plausible at all times. Therefore, in order to close the contour, Breadth First Search (BFS) algorithm is employed, whose pseudo code is presented in algorithm 2. In summary, the method receives as input the graph of all possible vertices and the starting point. Then, a queue and a set are created to process the explored nodes and keep track of the visited ones, respectively. The queue is initialized with the starting point, and through the iterations of the algorithm it gets filled with all the neighbors of the current node that is being explored. When a node is explored, it is popped out of the queue and the next one to be explored is selected following the FIFO (first in, first out) logic. The algorithm finishes when all nodes have been explored (the queue is empty). This FIFO logic ensures that vertices closer to the source vertex are visited before vertices farther away. Therefore, the algorithm is able to find the shortest path between any two points of a given graph, and that is the reason why it was selected for closing the contour of the NTMI. An example of an intermediate generated path from the BFS algorithm for NTMI is shown in figure 11

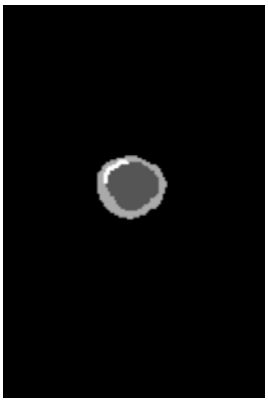


Figure 11: BFS generated path for closing contour of NTMI.

Creation of MI mask: Once the contour is closed, it is filled and the final MI is generated, as shown in image g from figure 10.

Generation of MI perturbation on the MRI: Following the rule described in the intensity priors section, the MI is generated as the mean value of the left ventricle cavity plus the prior knowledge value, which is a sampled value from a generated normal distribution whose mean is the average of the differences between the maximum intensity value of the MI minus the mean value of the left ventricle cavity across the entire pathological cases on EMIDEC dataset. As a final step, some blurring was applied with the goal of smoothing the intensity variations in the borders of the myocardium. In order to generate this effect, the blurring was applied on the area delimited by a dilated mask (for one iteration) of the myocardium using a 3x3 kernel. The final result of the algorithm is depicted in image h from figure 10.

Algorithm 2 BFS algorithm

```
function BFS(Graph, start)
    Create an empty queue Q
    Create a set to keep track of visited vertices
    Enqueue the start vertex into Q
    Add the start vertex to the visited set or array
    while Q is not empty do
        current_vertex ← dequeue from Q
        for neighbor in neighbors list of current_vertex do
            if neighbor is not visited then
                Add neighbor to the set of visited vertices
                Enqueue neighbor into Q
            end if
        end for
    end while
end function
```

3.3. Experiments

The experimentation protocols for this work where defined as follows:

- **Single modality baseline**, which corresponds to training the modified UNet for the target task.
- **Multi modality baseline**, which corresponds to training the DualUNet for the target task.
- **Evaluation of the proposed data augmentation algorithm**, which refers to measuring the impact of applying the data augmentation under two different scenarios: random and adaptive, which are explained in detail in the next section.
- **Further pushing the work towards a more robust pipeline:** The last part of the work consisted on using the nnUNet framework for the segmentation task in both single modality and multi modality (at input level) versions, since this framework has proven to be quite robust for different segmentation tasks regardless of the type of data used. Therefore, these experiments were performed to assess how such a framework behaves with our target segmentation task, and extract some useful insights on future directions of work to make the overall pipeline more robust.

3.4. Training details

3.4.1. 5-Fold cross validation

All the experiments were done using 5 fold-cross validation. Within each fold the data was split in the following way: 60% for training, 20% for validation and

20% for testing. The splits were made in such a way that the classes were equally distributed.

3.4.2. Application of the proposed augmentation algorithm

Random augmentation: This strategy refers to randomly enlarging the training dataset at different rates, e.g, 50%, 100% and 200% and evaluating its impact on the segmentation performance for both single modality and multi modality approaches.

Adaptive augmentation: In this case, the augmentation was applied by getting feedback of the model training, e.g, monitoring the cases in which it struggled the most. This criterion of "struggle" was defined as any case for which the myocardium Dice score was less than 0.8. The main idea behind this strategy is summarized in algorithm 3.

Algorithm 3 Adaptive data augmentation (ADA)

```
function ADA
    difficult_cases ← []
    for fold in range(5) do
        Train model
        Recover test set metrics in test_metrics array
        for DSC_myo in test_metrics do
            diff_cases_fold ← []
            if DSC_myo < 0.8 then
                Append DSC_myo to
                diff_cases_fold
            end if
        end for
        Append diff_cases_fold to difficult_cases
    end for
    for fold in range(5) do
        Augment difficult_cases[fold]
        Train model
    end for
end function
```

3.4.3. Auxiliary data augmentation techniques

On top of the proposed data augmentation algorithm, some other typical transformations were randomly applied to the data, namely, horizontal and vertical flips, rotations, random brightness contrast, random gamma (change the image contrast by raising its intensity values to the gamma power), and Contrast Limited Adaptive Histogram Equalization (CLAHE) with clip limit equal to 2 and grid size of (8,8).

3.4.4. Optimization

Throughout the experiments, an initial phase of hyperparameter tuning was always performed. The sweep included the following parameters:

- Loss function, which was varied between Dice Loss and Focal Dice Loss (a weighted average between the Focal and Dice loss).
- The selected optimizer was ADAM, proposed by Kingma and Ba (2014), with a learning rate varying randomly between 10^{-1} and 10^{-5}, and a weight decay fixed at 10^{-5}.
- Batch size, which varied between 4, 8 and 16.
- Early stopping, which was set with a patience of 20 epochs with a maximum number of 200 epochs.

The parameters for which the best results were consistently obtained in the baseline architectures were the following: Focal Dice Loss, ADAM optimizer with a learning rate of 3^{-4}, weight decay of 10^{-5} and a batch size of 8. A step scheduler was fixed during all training experiments, which decreased the learning rate every 5 epochs by a multiplicative factor of 0.95.

With regards to the nnUNet framework, the following experiments were made: the number of epochs was limited to 200, and the splits were redesigned according to the k-fold strategy explained in section 3.4.1. The loss function was also varied between Dice and Cross Entropy loss and TopK Dice loss. The best results were obtained with the first one.

3.4.5. Postprocessing

Once the segmentation masks were generated, the postprocessing for all experiments consisted of two simple steps: extraction of the largest area component and holes filling.

3.4.6. Implementation

The pipeline was implemented using Python 3.9.1, PyTorch 2.0 and CUDA 11.8 in Linux OS. For managing NIFTI files, the library nibabel 5.1.0 was used. For image processing tasks, the libraries OpenCV 4.7, Pillow 9.5.0 and scikit image 0.20.0 were used. The loss functions used for training the architectures were provided by MONAI 1.1.0 library.

All the training experiments were performed using NVIDIA Tesla V100 DGXS GPUs with 32 GB of RAM memory, provided by mésocentre de calcul de Franche-Comté, Besançon, France.

3.5. Metrics and statistic analysis of the results

3.5.1. Dice Score

The Dice score is one of the most commonly used metrics for evaluating the performance of image segmentation pipelines. It measures the overlap between predicted segmentation mask (PR) and the ground truth (GT). As presented in equation 1, the Dice score is mathematically defined as 2 times the intersection of both GT and PR (which refers to the effective overlap), divided by the total amount of pixels in both predicted mask and ground truth.

$$DSC = \frac{2 * |GT \cap PR|}{|GT| + |PR|} \quad (1)$$

This quantity ranges from 0 (no match) to 1 (perfect match). Some of its main advantages are that it is sensitive to true positive matches, robust to class imbalance, has an intuitive interpretation and it is differentiable, which makes it suitable to be used as a training loss function. Considering these benefits, the Dice score is selected as the main evaluation metric for all the experiments and astages of this work, which includes training optimization and models comparison.

3.5.2. Hausdorff distance

The Hausdorff distance is also a common metric used when evaluating image segmentation results. It typically quantifies the dissimilarity between two sets of points or shapes (contours), by measuring the maximum distance between any point in one set and its nearest point in the other set. In the context of image segmentation, this translates to assessing to which extent the predicted mask deviates from the ground truth with respect to shape and spatial arrangement. The smaller the Hausdorff distance, the better the alignment between the prediction and the ground truth.

Although this metric offers a different and useful perspective on the segmentation performance, the evaluation is focused at the contour level, its interpretation is less straightforward when the structures of interest are delimited by two contours (as in the case in the myocardium, which is bordered by epicardium and endocardium). Therefore, it is used in this work as an auxiliary local metric to provide a comprehensive overview of the pipeline performance exclusively in the final stage.

3.5.3. Statistical tests

For the segmentation results, the Dice scores are reported with standard deviations, and, in the final stage of methods comparison, the hausdorff distance is also reported. Similarly, to evaluate the effectiveness of applying our data augmentation algorithm to improve the segmentation of the target structure (DE myocardium), Wilcoxon signed-rank test, proposed by Wilcoxon (1992) is used in the final experiments. This method was selected since it allows to test the null hypothesis that two dependent samples (before and after the algorithm), which are not normally distributed (as in this case), come from the same distribution. If rejected, this means there is sufficient statistical evidence to conclude that the samples are different, and therefore that the algorithm had a meaningful impact on the metrics.

4. Results

4.1. Single modality

As explained in section 3, the first set of experiments were done under a single modality scenario i.e., only using image information from DE MRI. The segmentation architecture was a traditional UNet, adding batch normalization layers.

The first experiments consisted on evaluating the effect of our data augmentation algorithm in its fixed version, i.e, performing the augmentation at fixed rates. Tables 1 and 2 present the mean validation and test Dice score, respectively, with their corresponding standard deviation. "DA - percentage %", stands for data augmentation applied at a given rate, e.g, "DA - 100 %" means that for any given image, a synthetic one with the presence of myocardial infarction was also created.

By doing an overview of both tables, it can be noted that the generalization capability of the model is good, since the Dice score is either maintained or has a slight variation (either decrease or increase) between the validation and test stages. It can also be noticed that the data augmentation algorithm generates an increase in the myocardium Dice score. In terms of the validation set, it increases up to 100% DA, and then slightly decreases. In the case of the test set, the metric is directly proportional to the amount of data applied, i.e, the higher the amount of data augmentation, the higher the Dice score. It should be duly noted that, in the validation set, the standard deviation of the metric continuously decreases when more data is available (up to 100%). In the case of the test set, the standard deviation of the metric only decreases at 200% DA, which also happens to be the scenario under which the model performs the best, with a Dice score of 0.845 ± 0.052. With regards to the left ventricle cavity, the Dice score is higher in all cases where data augmentation was applied, with respect to the one where it was not. For this case, the best performance was also encountered at 200% DA, with a Dice score of 0.943 ± 0.034.

When doing a case-by-case analysis, it was found that most of the times the algorithm was able to notoriously improve the segmentation on the structure of interest (the myocardium), as shown by the signal areas of figure 12. However, the segmentation performance dropped slightly in a few cases, as shown in the signaled areas of figure 13.

The last experiment that was performed in a single modality setting was applying adaptive data augmentation. The results for this procedure are presented in table 3. It can be noted that for both validation and tests sets, the segmentation performance of both structures was improved when adding the data augmentation algorithm, and the standard deviation also decreased, although it was lower than the one obtained with data augmentation at a rate of 200%.

4.2. Multi modality

The same group of experiments done in single modality were also performed under the multi modality approach, which consisted on using DualUNet as a segmentation architecture, in which data from CINE and

Table 1: Validation set mean Dice score and standard deviation for single modality approach at different fixed data augmentation (DA) rates.

Modality	Structure	UNet	UNet - 50% DA	UNet - 100% DA	UNet - 200% DA
DE	CV	0.933 ± 0.059	0.939 ± 0.040	**0.939 ± 0.038**	0.936 ± 0.074
	MYO	0.826 ± 0.071	0.838 ± 0.056	**0.844 ± 0.052**	0.843 ± 0.055

Table 2: Test set mean Dice score and standard deviation for the single modality approach at different data augmentation (DA) rates.

Modality	Structure	UNet	UNet - 50% DA	UNet - 100% DA	UNet - 200% DA
DE	CV	0.936 ± 0.044	0.938 ± 0.041	0.938 ± 0.053	**0.943 ± 0.034**
	MYO	0.834 ± 0.056	0.838 ± 0.065	0.844 ± 0.058	**0.845 ± 0.052**

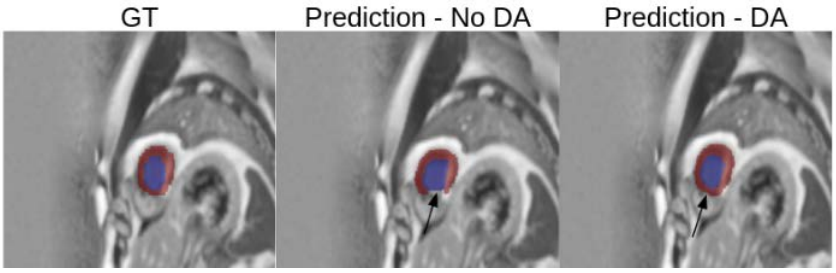


Figure 12: Visual example of a higher segmentation performance for the single-modality model after applying data augmentation. From left to right: ground truth, prediction of model without data augmentation, prediction of model with best data augmentation rate (200%). Left ventricle cavity and myocardium are presented in blue and red, respectively.

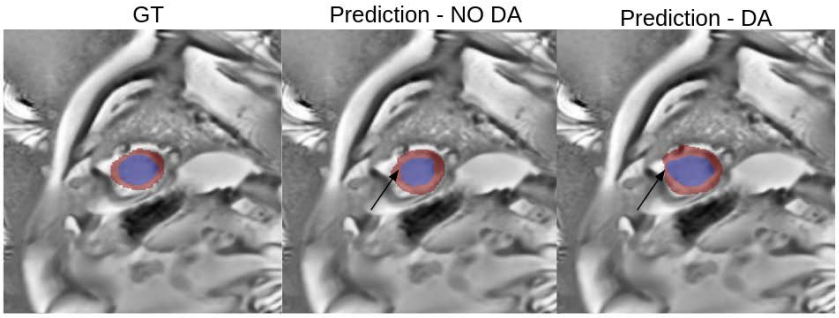


Figure 13: Example of a lower segmentation performance for the single-modality model after applying data augmentation. From left to right: ground truth, prediction of model without data augmentation, prediction of model with best data augmentation rate (200%). Left ventricle cavity and myocardium are presented in blue and red, respectively.

Table 3: Training and validation Dice scores for single modality segmentation trained without and with adaptive data augmentation (ADA).

Set	Modality	Structure	UNet	UNet - ADA
Validation	DE	CV	0.933 ± 0.059	**0.936 ± 0.051**
		MYO	0.826 ± 0.071	**0.841 ± 0.053**
Test	DE	CV	0.936 ± 0.044	**0.940 ± 0.035**
		MYO	0.834 ± 0.056	**0.843 ± 0.051**

DE modalities are fused at an intermediate step and the segmentation performance can be recovered for both, although bigger emphasis will be done in the modality of interest (DE MRI).

Table 4 and table 5 present the mean Dice score and its corresponding standard deviation for the validation and test sets, respectively. It can be noted that, similarly to the single-modality approach, the model generalization performance is good, as the segmentation metric remains the same or has slight variations (increase or

decrease) across all modalities and structures. Furthermore, it can be observed that, in terms of the segmentation performance for DE MRI, the model consistently gets higher Dice score and lower standard deviation as the data augmentation rate also increases. However, it reaches a plateau at 100%, after which the Dice score decreases and the standard deviation increases. The best performance in the test set is thus obtained when the DualUNet is trained on a dataset with 100% data augmentation, for which the Dice score is 0.856 ± 0.046.

Interestingly, it can be observed that, in the case of CINE modality, there is also an increase in the segmentation performance on the myocardium and the left ventricle cavity when the augmentation rate is 100%.

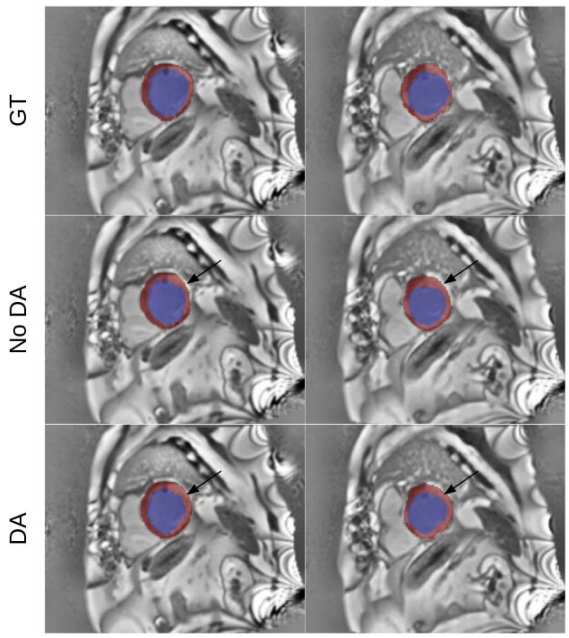


Figure 14: Examples of higher segmentation performance for the multi-modality model after applying data augmentation. From top to bottom : ground truth, prediction of model without data augmentation, prediction of model with the best data augmentation rate (100%). Left ventricle cavity and myocardium are presented in blue and red, respectively.

Throughout a case specific analysis, it was found that, in 89 out of 124 cases (71.77%), the Dice score was higher after applying the data augmentation algorithm. One example of this higher performance is presented in figure 14. When focusing on the areas selected by the arrows, it can be noted that the data augmentation algorithm indeed makes the model learn from different myocardial infarction shapes, thus making it capable of delineating the myocardium contour much more thoroughly, covering areas corresponding to myocardial infarction that were missed before applying the algorithm.

In the cases in which the Dice was lower, the visual difference was in general difficult to spot. Figure 15 presents examples of such cases, and it can be observed

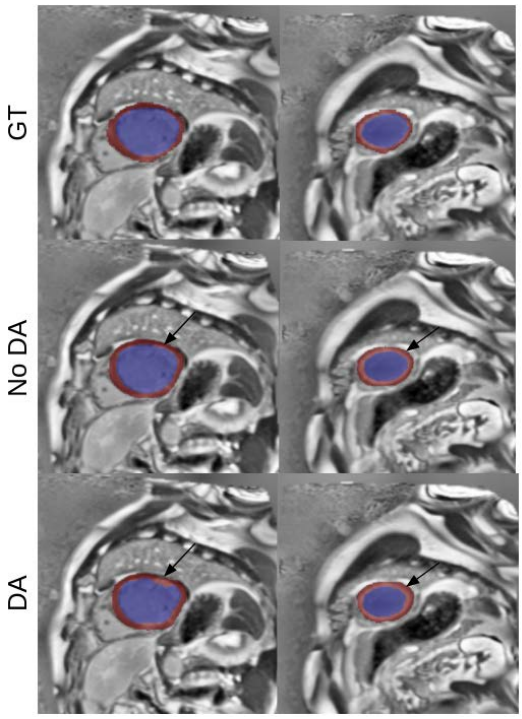


Figure 15: Examples of lower segmentation performance for the multi-modality model after applying data augmentation. From top to bottom : ground truth, prediction of model without data augmentation, prediction of model with the best data augmentation rate (100%).Left ventricle cavity and myocardium are presented in blue and red, respectively.

that in the area selected by the arrow, there is a slightly worse segmentation of the myocardium as it becomes thicker (taking some area that should have been delimited as left ventricle cavity). A second experiment was performed applying adaptive data augmentation based on the received training feedback from the model, and the results are reported in table 6. Once again, it can be easily seen that, for all structures of DE MRI modality, the Dice score is higher and the standard deviation is lower for both validation and test sets compared to no DA, but it is lower than the one obtained with the best fixed rate data augmentation strategy (100%).

4.3. Single vs Multi modality

Based on the information presented on tables 2 and 5, it can be seen that the performance of the best models for the myocardium segmentation under single and multi modality approaches are 0.845 ± 0.052 and 0.856 ± 0.040, respectively. Under Wilcoxon test, the null hypothesis that both results came from the same distribution was rejected ($p_{value} < 10^{-5}$), thus implying that the metrics difference is statistically significant and crossing information between CINE and DE modalities had a positive impact on the segmentation on DE myocardium. Figure 16 shows an example of the visual performance under the two scenarios, and it

Table 4: Validation set mean Dice score and standard deviation for multi modality approach at different data augmentation (DA) rates.

Modality	Structure	DualUNet	DualUNet - 50% DA	DualUNet - 100% DA	DualUNet - 200% DA
CINE	CV	0.948 ± 0.011	0.946 ± 0.012	0.947 ± 0.015	**0.949 ± 0.012**
	MYO	0.841 ± 0.029	0.841 ± 0.029	**0.849 ± 0.027**	0.847 ± 0.029
DE	CV	0.942 ± 0.028	0.942 ± 0.028	0.944 ± 0.029	**0.945 ± 0.027**
	MYO	0.843 ± 0.045	0.848 ± 0.044	0.852 ± 0.045	**0.853 ± 0.038**

Table 5: Test set mean Dice score and standard deviation for multi modality approach at different augmentation (DA) rates.

Modality	Structure	DualUNet	DualUNet - 50% DA	DualUNet - 100% DA	DualUNet - 200% DA
CINE	CV	0.946 ± 0.021	0.947 ± 0.012	**0.948 ± 0.013**	0.947 ± 0.012
	MYO	0.837 ± 0.037	0.841 ± 0.033	**0.848 ± 0.031**	0.846 ± 0.028
DE	CV	0.942 ± 0.032	0.942 ± 0.027	0.945 ± 0.028	**0.945 ± 0.027**
	MYO	0.845 ± 0.044	0.847 ± 0.046	**0.856 ± 0.040**	0.852 ± 0.042

Table 6: Training and validation Dice scores for multi modality segmentation trained without and with adaptive data augmentation (ADA).

Set	Modality	Structure	DualUNet	DualUNet - ADA
Validation	CINE	CV	0.948 ± 0.011	**0.948 ± 0.012**
		MYO	0.841 ± 0.029	**0.844 ± 0.029**
	DE	CV	0.942 ± 0.028	**0.945 ± 0.027**
		MYO	0.843 ± 0.045	**0.851 ± 0.040**
Test	CINE	CV	0.946 ± 0.021	**0.949 ± 0.011**
		MYO	0.837 ± 0.037	**0.848 ± 0.029**
	DE	CV	0.942 ± 0.032	**0.944 ± 0.027**
		MYO	0.845 ± 0.044	**0.853 ± 0.039**

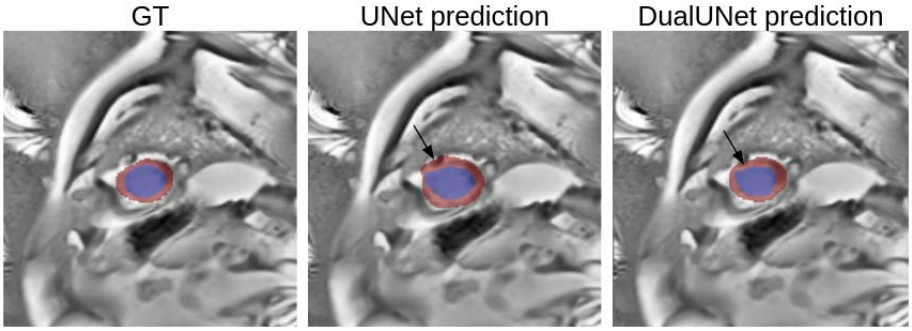


Figure 16: Example of segmentation performance under single and multi modality approaches. Left ventricle cavity and myocardium are presented in blue and red, respectively.

is clearly observed that in the multi-modality approach a more consistent segmentation of the myocardium is produced, thus this pipeline (DualUNet trained at 100% enlarged dataset, hereinafter called "DualUNet100"), is deemed to be the overall best performing method for the segmentation of the myocardium in DE MRI.

4.4. Best model further results

The final experiment in this stage consisted on comparing DualUNet vs DualUNet100. Table 7 shows the segmentation results for these two methods in terms of both metrics: Dice score and Hausdorff distance. It can be seen that, for the validation and test sets, the myocardium segmentation Dice score is consistently higher after applying the data augmentation strategy, and the standard deviation is also smaller. This difference was proven to be statistically significant, under a Wilcoxon test with $p_{value} < 10^{-5}$. In terms of the Hausdorff distance, since the region of interest is bordered by two contours, namely, epicardium (CV) and endo-

Table 7: Training and validation mean Dice score and Hausdorff distance of the best standalone model (DualUNet) and after applying the best data augmentation (DA) strategy (DualUNet100). DSC: Dice score.

Set	Modality	Structure	DualUNet DSC	DualUNet (100 % DA) DSC	DualUNet HD (mm)	DualUNet (100 % DA) HD (mm)
Validation	CINE	CV	**0.948 ± 0.011**	0.947 ± 0.015	**6.36 ± 3.54**	6.39 ± 4.65
		MYO	0.841 ± 0.029	**0.849 ± 0.027**	6.51 ± 1.71	**6.01 ± 1.41**
	DE	CV	0.942 ± 0.028	**0.944 ± 0.029**	**5.67 ± 1.93**	6.03 ± 2.58
		MYO	0.843 ± 0.045	**0.852 ± 0.045**	6.09 ± 2.03	**6.02 ± 1.85**
Test	CINE	CV	0.946 ± 0.021	**0.948 ± 0.013**	6.39 ± 4.64	**5.73 ± 2.12**
		MYO	0.837 ± 0.037	**0.848 ± 0.031**	**6.49 ± 1.78**	6.51 ± 1.71
	DE	CV	0.942 ± 0.032	**0.945 ± 0.028**	**5.73 ± 2.12**	5.77 ± 2.23
		MYO	0.845 ± 0.044	**0.856 ± 0.040**	6.08 ± 1.75	**5.90 ± 1.93**

cardium (MYO), they both have to be taken into account. It can be observed that, in the test set, the Hausdorff distance is slightly lower for the endocardium and slightly higher for the epicardium, although the increase difference is higher than the decrease one, this might indicate that, overall the segmentation performance at the contour level is approximately maintained.

With the goal of performing a more extensive study, the segmentation performance of cases corresponding to the VIA class (myocardial infarction) was analyzed. It was hypothesized that, if the proposed data augmentation was working as expected, it should be able to reduce the outliers for the cases classified as myocardial infarction, since the synthetic data that is generated aims to provide the models with more samples of this particular condition to improve the segmentation of the myocardium.

Figure 17 shows the distribution of the myocardium Dice score before and after applying the data augmentation algorithm. As expected, the overall distribution of the metric went up, and two of the three outliers were eliminated. Although the worst one was not removed, its score improved significantly.

The DualUNet architecture that was used for the multi-modality approach was investigated in the work of Ouadah et al. (2022), and they achieved a mean myocardium test set Dice score of 0.81 ± 0.06. On the other hand, in this study we obtained a mean myocardium test set Dice score of 0.845 ± 0.044 and 0.856 ± 0.040 on the best model without and with our data augmentation algorithm, respectively, which is in both cases substantially higher and more concentrated around the mean value, which gives evidence of an increase of robustness of the overall pipeline.

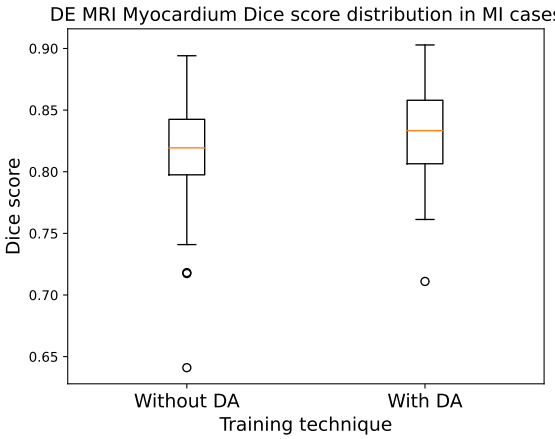


Figure 17: Myocardium Dice Score distribution on myocardial infarction cases before and after applying data augmentation.

4.5. nnUNet framework results

With the goal of getting some insights on how to further improve the results on the myocardium segmentation and construct a more robust pipeline, in the last part of the work some experiments were done to assess the performance of the nnUNet framework for this particular task in both single modality and multi modality, which is done at the input level, i.e., the CINE modality is added as an extra channel of the target one (DE). Table 8 presents a summary of the mean Dice score and Hausdorff distance for all structures on both validation and test sets. Based on the results from table 8, it can be seen that, in the test phase, the results for multi modality are slightly worse than those of single modality, specially in terms of the Hausdorff distance.

Also, when compared to DualUNet100, the results of nnUNet are better in both metrics (higher Dice score and lower Hausdorff distance), although the Dice score standard deviation is higher, which could give some signs of higher variability. It should be duly noted that the nnUNet was trained without data augmentation, i.e., our data augmentation algorithm was not applied due to time restrictions.

Table 8: nnUNet framework results. DSC: Dice score.

Set	Modality	Structure	nnUNet DSC		nnUNet HD (mm)	
			Single modality	Multimodality	Single modality	Multi modality
Validation	DE	CV	0.949 ± 0.030	**0.949 ± 0.028**	3.65 ± 1.94	**3.61 ± 1.68**
		MYO	0.861 ± 0.051	**0.863 ± 0.045**	3.64 ± 1.19	**3.61 ± 1.24**
Test	DE	CV	**0.950 ± 0.029**	0.949 ± 0.028	**3.55 ± 1.82**	3.60 ± 1.59
		MYO	0.862 ± 0.054	**0.862 ± 0.045**	**3.58 ± 0.99**	3.60 ± 1.10

5. Discussion

In this study, we have proposed an image processing based data augmentation algorithm that, taking some prior knowledge extracted from an external dataset, is able to create synthetic infarction cases to allow deep learning models to further learn from this condition as there is generally scarcity of such type of data, and it was previously found that segmentation models struggled a lot when segmenting the myocardium in such cases.

The proposed data augmentation approach works under two possible scenarios: fixed, and adaptive. In the first one, the training dataset is enlarged at a given amount, whereas in the second, feedback from the model is obtained while training. This feedback consists on tracking the samples on which the segmentation performance was low, and then only creating synthetic MI on those particular cases.

The first set of experiments were performed under the single modality approach. This was done using a traditional UNet with batch normalization layers as segmentation model. Afterwards, the data augmentation algorithm was applied at fixed rates, namely, 50%, 100%, and 200%. Augmenting data at 100% means that for every case in the dataset, a new one with the presence of synthetic MI was also created. According to the results presented in tables 1 and 2, it could be noticed that the segmentation performance on the myocardium was higher as more artificial data was available, the best result being obtained at a rate of 200%. As evidenced in table 3, the adaptive data augmentation strategy offered an increase of the Dice score of the myocardium, but this was lesser than the increase generated with the 200% strategy.

In a similar fashion, the results of the multi-modality approach reported on tables 4 and 5 show that the myocardium test Dice score consistently improved when increasing the percentage of data augmentation up to 100%, where it reached a plateau state and decreased with higher augmentation rates. Once again, the results were also improved with the adaptive data augmentation strategy, but to a smaller extent when comparing it to the 100% dataset enlargement.

In both single and multi modality scenarios it was possible to notice that the models have a good generalization capability, as the mean test Dice score across all 5 folds is generally maintained or has slight changes with respect to the validation one. Furthermore, the visual results shown in figures 12 and 14 demonstrate that adding the data augmentation algorithm had an effective impact on allowing the segmentation models to better delineate the myocardium area, completing the contours as in the case of single modality and providing more accurate segmentations that include the myocardial infarction areas that had been previously missed in the multi-modality case.

Nonetheless, there are a few cases where the segmentation performance is lower after applying the algorithm, such as the cases reported in figures 13 and 15, where the myocardium area gets thicker than it should (thus classifying left ventricle cavity pixels as part of the myocardium). Further investigation should be done in order to limit this minority cases. Future work should be focused on tuning the most optimal range of values for these parameters to make the learning of myocardial infarction more robust, which also includes investigating more possible intensity patterns when creating the myocardial infarction on DE MRI.

When comparing single vs multi-modality approach, it was found that this latter consistently outperformed the first in both structures (myocardium and left ventricle cavity). Figure 16 proved that, in general, using information from CINE modality has a positive impact on the segmentation of the myocardium.

We obtained significantly better test results when compared to the work of Ouadah et al. (2022), where the same dataset and architecture was used. This gives clear evidence of the increase of robustness of the segmentation achievable with the addition of the data augmentation algorithm.

As an added value of our work, we have extended the study by assessing the segmentation performance of new deep learning frameworks such as nnUNet in both single modality and multi modality approaches in our dataset. After doing some modifications on the framework, such as the design of cross-validation training splits and loss function to use, it was found that the performance of the multi modality approach, which is done at the input level (adding the second modality as a channel of the target one), was slightly worse than the one of single modality. This is in agreement with the conclusions of the work of Ouadah et al. (2022), where they found that an input fusion UNet had a worse segmentation performance than a single modality UNet.

The results reported in section 4.5 showcase the high potentials of the nnUNet for segmentation tasks. As shown in table 8, the method got a higher Dice score for the myocardium than DualUNet with data augmentation. Although the difference is not much, and the nnUNet results have higher variability, this is quite an interesting finding since it opens up the doors for future work within this field of research, which should be directed towards adding our data augmentation algorithm into the nnUNet framework. Although this step is not trivial, we hypothesize that it would generate quite a robust pipeline for the segmentation of the myocardium of DE MRI, since we proved that the data augmentation algorithm improved the performance and robustness of the selected architectures during this work, and since the nnUNet is already packed with powerful techniques for automatic data preprocessing, feeding it with a more diverse range of data where myocardial infarction is present will most likely have a positive effect.

The addition of the data augmentation could be done at a fixed rate, but it could even be more interesting to further explore the adaptive version, as it was proven that this strategy managed to increase the segmentation results and uses less data, thus having a lower computational cost. As noted in the nnUNet experiments, the multi modality approach (which is done at the input level), was slightly worse than the single modality one. Therefore, as a first step, the adding of the data augmentation should be done in a single modality setting. Nonetheless, there is also room for improvement in a multi modality approach, since the nnUNet framework could be further modified to support data fusion at an intermediate step (as in DualUNet), and given the fact that both architectures are UNet based, it is likely that the effect of the data augmentation will be as positive as it was proven to be on DualUNet.

6. Conclusions

In this work, we have proposed a data augmentation algorithm where we create synthetic myocardial infarction shapes to allow deep learning models to further learn from pathological cases and improve the segmentation of the myocardium on Delayed Enhacement MRI. We proved that the method made the overall pipeline more robust, obtaining higher scores and less variability when compared to a previous work in this application. We have shown that the algorithm can be applied under two different strategies, and both of them managed to improve the segmentation results. We have also done some additional experiments assessing the performance of newer frameworks such as nnUNet, and our findings suggest some future work in this regard, such as adding our data augmentation algorithm within the framework, either under the fixed or adaptive scenario, or as a hybrid of both, and also adding the DualUNet fusion scheme. Some other lines of future work include further exploring the tuning of the parameters of our data augmentation algorithm, and extending the study in 3D.

Acknowledgments

I would like to express my deepest gratitude to my supervisors, Dr. Alain Lalande and Dr. Sarah Leclerc, for their continuous guidance, support and advice throughout this work.

I also want to thank all professors and the coordination team of the MAIA program for making these past two years an incredible and memorable experience of academic and personal growth.

I am immensely grateful to my parents and my brother for always believing in me. Your endless love, support and encouragement despite distance has been my biggest source of inspiration and motivation.

References

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al., 2018. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE transactions on medical imaging 37, 2514–2525.

Bresenham, J.E., 1965. Algorithm for computer control of a digital plotter. IBM Systems journal 4, 25–30.

Cerqueira, M.D., Weissman, N.J., Dilsizian, V., Jacobs, A.K., Kaul, S., Laskey, W.K., Pennell, D.J., Rumberger, J.A., Ryan, T., et al., 2002. Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart: a statement for healthcare professionals from the cardiac imaging committee of the council on clinical cardiology of the american heart association. Circulation 105, 539–542.

Chen, C., Qin, C., Ouyang, C., Li, Z., Wang, S., Qiu, H., Chen, L., Tarroni, G., Bai, W., Rueckert, D., 2022. Enhancing mr image segmentation with realistic adversarial data augmentation. Medical Image Analysis 82, 102597.

Ciofolo, C., Fradkin, M., Mory, B., Hautvast, G., Breeuwer, M., 2008. Automatic myocardium segmentation in late-enhancement mri, in: 2008 5th IEEE International Symposium on Biomedical Imaging: from nano to macro, IEEE. pp. 225–228.

Dikici, E., O'Donnell, T., Setser, R., White, R.D., 2004. Quantification of delayed enhancement mr images, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2004: 7th International Conference, Saint-Malo, France, September 26-29, 2004. Proceedings, Part I 7, Springer. pp. 250–257.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020. Generative adversarial networks. Communications of the ACM 63, 139–144.

Huellebrand, M., Ivantsits, M., Zhang, H., Kohlmann, P., Kuhnigk, J.M., Kuehne, T., Schönberg, S., Hennemuth, A., 2021. Comparison of a hybrid mixture model and a cnn for the segmentation of myocardial pathologies in delayed enhancement mri, in: Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers 11, Springer. pp. 319–327.

Isensee, F., Jaeger, P.F., Full, P.M., Wolf, I., Engelhardt, S., Maier-Hein, K.H., 2018. Automatic cardiac disease assessment on cine-mri via time-series segmentation and domain specific features, in: Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges: 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec

City, Canada, September 10-14, 2017, Revised Selected Papers 8, Springer. pp. 120–129.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods 18, 203–211.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Lalande, A., Chen, Z., Decourselle, T., Qayyum, A., Pommier, T., Lorgis, L., de la Rosa, E., Cochet, A., Cottin, Y., Ginhac, D., et al., 2020. Emidec: a database usable for the automatic evaluation of myocardial infarction from delayed-enhancement cardiac mri. Data 5, 89.

Lalande, A., Chen, Z., Pommier, T., Decourselle, T., Qayyum, A., Salomon, M., Ginhac, D., Skandarani, Y., Boucher, A., Brahim, K., et al., 2022. Deep learning methods for automatic evaluation of delayed enhancement-mri. the results of the emidec challenge. Medical Image Analysis 79, 102428.

Lin, A., Wu, J., Yang, X., 2020. A data augmentation approach to train fully convolutional networks for left ventricle segmentation. Magnetic Resonance Imaging 66, 152–164.

Mendis, S., Thygesen, K., Kuulasmaa, K., Giampaoli, S., Mähönen, M., Ngu Blackett, K., Lisheng, L., group on behalf of the participating experts of the WHO consultation for revision of WHO definition of myocardial infarction, W., 2011. World health organization definition of myocardial infarction: 2008–09 revision. International journal of epidemiology 40, 139–146.

Ouadah, C., Lalande, A., Leclerc, S., 2022. Fusion strategies for multi-modal left ventricle segmentation. MAIA MSc Thesis .

Petitjean, C., Dacher, J.N., 2011. A review of segmentation methods in short axis cardiac mr images. Medical image analysis 15, 169–184.

Rohé, M.M., Sermesant, M., Pennec, X., 2018. Automatic multi-atlas segmentation of myocardium with svf-net, in: Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges: 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers 8, Springer. pp. 170–177.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer. pp. 234–241.

Skandarani, Y., Painchaud, N., Jodoin, P.M., Lalande, A., 2020. On the effectiveness of gan generated cardiac mris for segmentation. arXiv preprint arXiv:2005.09026 .

Thomson, L.E., Kim, R.J., Judd, R.M., 2004. Magnetic resonance imaging for the assessment of myocardial viability. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine 19, 771–788.

Thygesen, K., Alpert, J.S., Jaffe, A.S., Simoons, M.L., Chaitman, B.R., White, H.D., 2012. Third universal definition of myocardial infarction. Journal of the American College of Cardiology 60, 1581–1598.

Tziritas, G., Grinias, E., 2017. Fast fully-automatic localization of left ventricle and myocardium in mri using mrf model optimization, substructures tracking and b-spline smoothing, in: International Workshop on Statistical Atlases and Computational Models of the Heart, pp. 91–100.

Wilcoxon, F., 1992. Individual comparisons by ranking methods. Springer.

Wolterink, J.M., Leiner, T., Viergever, M.A., Išgum, I., 2018. Automatic segmentation and disease classification using cardiac cine mr images, in: Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges: 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers 8, Springer. pp. 101–110.

Xue, Y., Farhat, F.G., Boukrina, O., Barrett, A., Binder, J.R., Roshan, U.W., Graves, W.W., 2020. A multi-path 2.5 dimensional convolutional neural network system for segmenting stroke lesions in brain mri images. NeuroImage: Clinical 25, 102118.

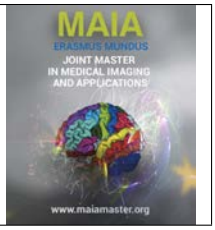
Yang, S., Wang, X., 2021. A hybrid network for automatic myocardial infarction segmentation in delayed enhancement-mri, in: Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers 11, Springer. pp. 351–358.

Zhang, Y., 2021. Cascaded convolutional neural network for automatic myocardial infarction segmentation from delayed-enhancement cardiac mri, in: Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers 11, Springer. pp. 328–333.

Zotti, C., Luo, Z., Lalande, A., Jodoin, P.M., 2018. Convolutional neural network with shape prior applied to cardiac mri segmentation. IEEE journal of biomedical and health informatics 23, 1119–1128.

Medical Imaging and Applications

Master Thesis, June 2023



Large Intestine 3D Shape Refinement Using Conditional Latent Point Diffusion Models

Kaouther Mouheb, Joseph Y. Lo Ph.D.

Center for Virtual Imaging Trials, Department of Radiology, Duke University School of Medicine, Durham, NC, USA.

Abstract

Accurate 3D modeling of human organs plays a crucial role in building computational phantoms for virtual imaging trials. However, generating anatomically plausible reconstructions of organ surfaces from computed tomography (CT) scans remains challenging for many structures in the human body. This challenge is particularly evident when dealing with the large intestine. In this study, we leverage recent advancements in geometric deep learning (DL) and denoising diffusion probabilistic models (DDPMs) to refine the segmentation results of the large intestine. We begin by representing the organ as point clouds sampled from the surface of the 3D segmentation mask. Subsequently, we employ a hierarchical variational autoencoder (VAE) to obtain global and local latent representations of the organ's shape. We train two conditional denoising diffusion models in the hierarchical latent space to perform shape refinement. To further enhance our method, we incorporate a state-of-the-art surface reconstruction model, allowing us to generate smooth meshes from the obtained complete point clouds. Experimental results demonstrate the effectiveness of our approach in capturing both the global distribution of the organ's shape and its fine details. Our complete refinement pipeline demonstrates remarkable enhancements in surface representation compared to the initial segmentation, reducing the Chamfer distance by 70%, the Hausdorff distance by 32%, and the Earth Mover's distance by 6%. By combining geometric DL, DDPMs, and advanced surface reconstruction techniques, our proposed method offers a promising solution for accurately modeling the large intestine's surface and can easily be extended to other anatomical structures.

Keywords: Large intestine modeling, 3D shape refinement, Computational phantom

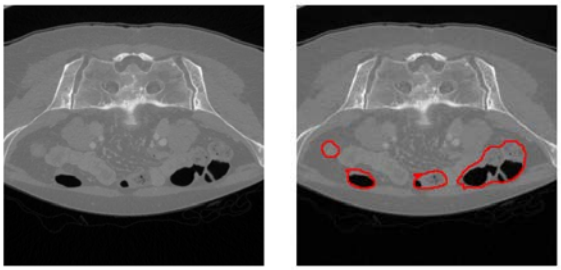
1. Introduction

In the field of virtual imaging trials, developing realistic digital phantoms is crucial because they serve as the virtual patients on which the simulated studies can be conducted. Combined with imaging and diagnostic simulation tools, they provide researchers with the ability to conduct iterative experiments involving limitless parameter combinations, alleviating concerns about potential side effects such as excessive radiation exposure. (Segars et al. (2013)). Computational phantoms have been widely used for many applications such as radiation dosimetry (Hesterman et al. (2017)) and radiotherapy (Wang et al. (2016)). In addition, compared to clinically generated data, these models can provide pixel-level ground truth, can be used to generate unlimited diverse data, and suffer fewer privacy and regulatory issues. Therefore, they are very useful to train and evaluate deep learning algorithms (Chen et al. (2022)). A typical pipeline for building computerized patient models includes collecting imaging data of real patients, segmenting the selected organs, and finally converting the generated masks to an easily deformable mathematical representation such as polygon meshes (Segars et al. (2010)). The segmentation step is the most critical part of this process and its results significantlyaffect the quality of the generated phantoms. For this purpose, early studies predominantly employed either manual contouring (Segars et al. (2010)) or conventional image processing-based segmentation algorithms, such as thresholding (Lee et al. (2007)), whereas recent approaches are shifting towards deep learning-based segmentation techniques (Dahal et al. (2023)).

Manual segmentation of numerous anatomical structures from 3D CT volumes is both subjective and time-consuming making its application to multiple patients impractical. Classical image processing methods often require manual refinement or hand-crafted post-processing pipelines that are organ-specific, rendering them inconvenient for constructing whole-body human phantoms with over a hundred structures. The advent of deep learning-based segmentation algorithms, coupled with the growing availability of radiological images, has led to the emergence of various deep learning-based medical image segmentation approaches, some of which can segment a considerably large number of organs such as TotalSegmentator (Wasserthal et al. (2022)). Even though these methods significantly outperform their traditional counterparts, the challenge of medical image segmentation is still far from being solved (Cerrolaza et al. (2019)). Current DL algorithms are based on volumetric CNN models such as U-Nets (Ronneberger et al. (2015)) which can accurately recover organ volumes but struggle in obtaining accurate surface representations due to multiple factors. These models do not take shape constraints into consideration and they output discrete voxel grids which lead to discretization effects when converted to surface representations, resulting in anatomically inaccurate shapes (Yang et al. (2022), Raju et al. (2022)).

Some organs suffer from these issues more than others. The large intestine is one of the hardest organs to accurately reconstruct from CT scans due to its complex characteristics: the low contrast between the soft tissue of the intestine and its surroundings and the high heterogeneity of its filling, the high variability of the organ's shape, size, and appearance between different patients, the proximity to other abdominal organs with low-contrast boundaries, and the filling status of the organ that can highly affect its shape (Wang et al. (2022)). Under these conditions, U-Nets tend to generate incomplete segmentation masks that often incorporate sections of adjacent organs such as the small intestine, or organs that exhibit similar features such as the presence of air in the stomach.

While there has been limited research conducted on the topic of 3D shape completion and refinement in medical imaging, it is a widely explored problem in other domains including autonomous driving, robotics, and manufacturing. In these fields, Li-DAR sensors are commonly used to detect shapes, which often result in sparse, noisy, and partial shapes represented as point clouds. Consequently, it becomes crucial to develop tools for shape completion, denoising, and refinement. Notably, a significant breakthrough has been achieved in processing such data through the introduction of PointNets, which are deep networks specifically designed for point cloud data (Qi et al. (2017)). Additionally, denoising diffusion probabilistic models (Ho et al. (2020)) have demonstrated remarkable outcomes



(a) (b)

Figure 1: Example CT slice (a) and the corresponding large intestine contours (b) illustrating the heterogeneity of filling and the indistinct boundaries of the organ.

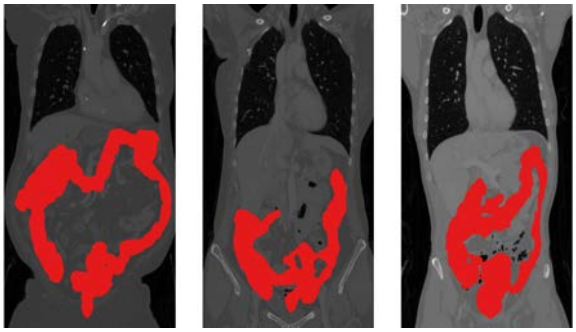


Figure 2: Coronal projections of the large intestine across different patients. Note the contrasting shape, size, and appearance with the standard textbook representation.

in completing missing data, whether it pertains to 1D signals like speech, 2D representations like images, or 3D structures such as object shapes. This research aims to leverage the advancements made in the aforementioned areas of deep learning to address the issue of inaccurate surface reconstructions of the large intestine resulting from a volumetric segmentation model. In particular, we represent the shapes as point clouds and train latent denoising diffusion models with PointNet-based backbones, conditioned on the partial shapes of the organ, to generate complete shapes as outputs. We also benefit from a modern surface reconstruction method to represent the final outputs as polygon meshes, which is a preferred representation in computational phantoms.

2. State of The Art

2.1. Multi-organ Segmentation from CT Scans: The Large Intestine Problem

In recent years, researchers have conducted numerous studies addressing the issue of multi-organ segmentation from CT scans, including the large intestine. Some notable studies include those by Liu et al. (2020) and Weston et al. (2020). More recently, Wasserthal et al. (2022) used a large dataset of 1024 CT scans along with the state-of-the-art medical image segmentation framework nn-Unet (Isensee et al. (2021)) to segment 104

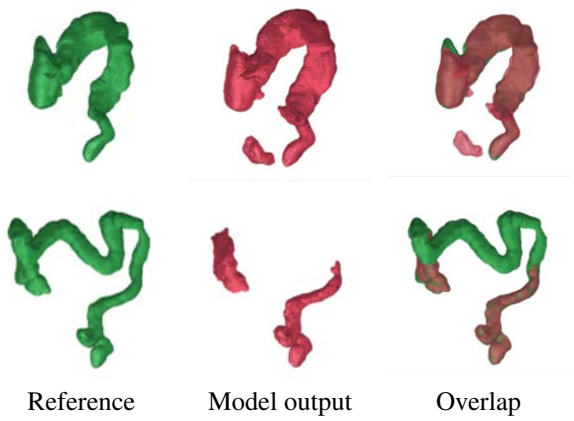


Figure 3: Examples of erroneous TotalSegmentator results with the corresponding references annotated by a physician. The segmentation failures include false positives included from other organs (upper row) and missing components (lower row).

anatomical structures. However, these studies primarily rely on volumetric CNNs, which do not take geometric features into consideration. Consequently, they struggle to accurately reconstruct the surface of the large intestine, even if the volume is successfully recovered. These failures are directly related to the previously mentioned complex characteristics of this organ's surface and filling as illustrated in Fig. 1 and Fig. 2. An inspection of TotalSegmentator's results on new CT scans revealed that the segmentation masks are often disconnected, partial, or contain portions of other structures (Fig. 3).

Oda et al. (2021) proposed a solution by converting segmented intestinal regions into a graph representation and then reconstructing the regions using a graph-based algorithm. Although qualitative evaluation demonstrated improved surface reconstruction, quantitative results were limited, and the method lacks the adaptability to other organs. Wang et al. (2022) propose BowelNet, a two-stage segmentation approach for the intestinal region. The first stage uses a CNN trained with fully labeled data to segment the bowel region, whereas the second stage employs another CNN trained with both partially and fully labeled data to refine the segmentation by incorporating geometric information. Instead of injecting shape information into a volumetric model, our study focuses on applying geometric models, designed specifically for 3D shape learning, to address the problem of inaccurate large intestine surface reconstruction.

2.2. 3D Shape Representations in Medical Imaging

Unlike 2D images, which are commonly represented as 2D matrices, 3D objects lack an established method for representation, with various approaches used in the literature indicating the absence of a singular standard in this domain. Here we describe the most widely used 3D shape representations in medical imaging.

2.2.1. Voxel Grids

Voxel-based volumes are the most common representation in medical imaging. Objects are annotated on discrete grids stored in 3D matrices, with each voxel indicating whether the corresponding volume is occupied by an object or not. This expands upon the 2D pixel grid representation of 2D images and enables the use of CNNs with 3D convolutions. The BowelNet model mentioned earlier employed voxel grids to represent the organs of the intestinal region.

2.2.2. Meshes

This involves representing a 3D object as a collection of interconnected polygons, typically triangles. For instance, Kong et al. (2021) propose a novel approach that uses a graph convolutional neural network to predict deformation on mesh vertices from a pre-defined mesh template and reconstruct multiple anatomical structures in a 3D image volume.

2.2.3. Point Clouds

In this setting, a 3D object is represented as a set of unordered points in 3D space, sampled from its outer surface, with each point having a set of associated features such as its coordinates in space or the components of its surface normal. Cai et al. (2019) introduce an end-to-end shape learning architecture that generates organ point clouds starting from deep features. Then the model is optimized to further refine the generated clouds using a novel adversarial shape learning loss. This model is trained alongside a CNN-based segmentation model for multi-task learning.

2.2.4. Implicit Representations

Implicit functions provide alternative representations for shapes, such as signed distance functions (SDFs) and occupancy maps. SDFs map points in 3D space to their closest distance to the object's surface, where a positive distance is used for points outside the object of interest and a negative distance is used for those inside it. Raju et al. (2022) utilize SDFs to model organ shapes, proposing a deep implicit statistical shape model for smooth surface generation. Occupancy maps indicate the likelihood of object occupancy in 3D grids. Yang et al. (2022) introduce ImplicitAtlas, which learns multiple organ templates of an organ that undergo non-rigid deformations and outputs occupancy maps.

Because the objective of this work is to train probabilistic generative models, point clouds represent an ideal representation (Zeng et al. (2022)). Compared to their counterparts, they are more compact and flexible, they can capture complex 3D shapes and model the distribution of geometric properties with fewer memory and computational resources. Therefore, we use point clouds as a 3D shape representation in this work.

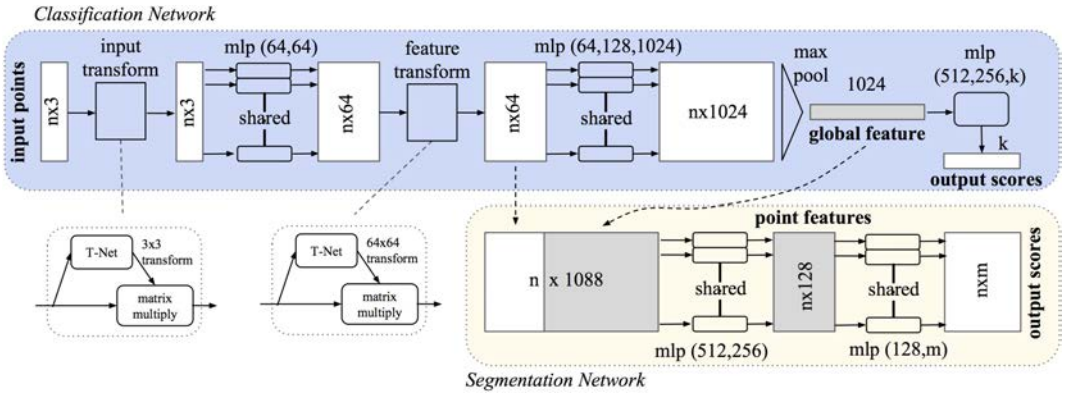


Figure 4: PointNet architecture. The classification network takes n points as input and applies input and feature transformations before aggregating point features using max pooling. The output is a classification score for m classes. The segmentation network is an extension of the classification network which concatenates global and local features before outputting per-point scores. Reproduced from Qi et al. (2017).

2.3. Deep Learning on Point Clouds

2.3.1. PointNets

Early deep learning techniques, like CNNs and Recurrent Neural Networks (RNNs), have been applied to point clouds for tasks like recognition and completion (Liu et al. (2019a), Fan and Yang (2019)). However, these methods require regularizing the point clouds which can be computationally demanding. To address this, Qi et al. (2017a) introduced PointNet, a deep architecture specifically designed for unordered data such as point clouds. PointNet processes each point independently and aggregates the learned features to create a global feature vector representing the entire cloud. The architecture of the vanilla PointNet model is depicted in Fig 4. PointNet++ (Qi et al. (2017b)) improved upon this architecture by using hierarchical feature learning, applying PointNets to small clusters of points to learn local features, and then feeding them to other PointNets for learning higher-level features from larger clusters. The network incorporates farthest-point sampling to select informative points and a multi-scale grouping strategy to aggregate features across different scales.

PointNets are memory-efficient but require structuring the data for neighbor querying and dynamic kernel computation due to irregular point distances. Liu et al. (2019b) introduced Point-Voxel CNN (PVCNN) to address this by combining PointNets' memory efficiency with the regularity of voxel-based models. PVCNN is based on a new primitive operation called Point-Voxel Convolution (PVConv), which includes a point branch for learning local features using MLPs and a voxel branch for learning coarse features using 3D convolutions.

Deep models for point cloud processing have been explored in medical imaging. For instance, Balsiger et al. (2019) propose a point cloud-based approach to refine peripheral nervous system segmentations. The method employs a CNN to segment the nervous system and extract image information from an MRN volume. Then point clouds are extracted from the mask's surface and the xyz-coordinates of each point are concatenated with its deep image features. Later, a PointCNN (Li et al. (2018)) is used to classify each point as foreground or background, refining the segmentation by eliminating the noise. However, this method does not address false negatives missed by the segmentation network. In contrast, we propose to use generative models that can generate new points to fill the missing parts in the organ's shape representation.

2.3.2. Generative Point-Diffusion Models

Denoising diffusion probabilistic models (DDPMs) (Ho et al. (2020)) have shown great potential in point cloud generation. They are designed to model the probability distribution of the 3D shapes in the dataset. During training, noise is gradually introduced to a clean shape and the model learns to predict the noise added at each step, minimizing the difference between the predicted and actual noise levels. At inference, the model denoises a diffused shape iteratively until a clean version is obtained (Luo and Hu (2021), Nichol et al. (2022)). Recent studies suggest performing the generation process in a latent space (Vahdat et al. (2021), Chou et al. (2022)). The latent Point Diffusion Model (LION) by Zeng et al. (2022) is a proposal for a hierarchical point diffusion model. LION utilizes a variational autoencoder (VAE) with a hierarchical latent space, combining global and local representations. By training diffusion models in this smoother latent space, the model can better capture shape distributions and generate higher-quality shapes. LION has achieved state-of-the-art performance on various benchmarks and can be extended to other applications such as text-to-shape generation.

2.3.3. Point-Diffusion Models for Shape Refinement

Although PointNet-based diffusion models were widely explored for point cloud generation, little work

has been done for point cloud completion and refinement. Zhou et al. (2021) propose a unified PVCNN-based architecture named Point-Voxel Diffusion (PVD) for both generation and completion. In this setting, shape completion is formulated as a conditional generation problem where the partial input shape is fixed and only the remaining points are diffused. In this case, the DDPM learns to model the missing parts in the data. Lyu et al. (2021) introduce a Point Diffusion-Refinement (PDR) approach that consists of a Conditional Generation Network (CGNet) and a ReFinement Network (RFNet). CGNet is a conditional DDPM that produces a coarse complete point cloud guided by its partial input, while RFNet densifies the generated point cloud for further quality improvement. The two networks use a dual-path architecture based on PointNet++ for efficient feature extraction from partial shapes and accurate manipulation of 3D point locations.

Friedrich et al. (2023) use PVD's shape completion method to generate 3D skull implants. The model is trained to generate a complete skull shape starting from a defective one. During the diffusion process, only the points belonging to the implant are modified, while those belonging to the defective structure remain unchanged. Finally, the defective shape is subtracted from the generated complete one to obtain the shape of the implant. In contrast to this work, our focus is on a deformable internal organ characterized by more intricate shapes and higher diversity across patients rather than the skull, which is a rigid anatomical structure with numerous fixed landmarks. Additionally, while this work only aims to fill the missing parts in the organ, we are also interested in reducing the false positives included from other organs.

2.4. Contributions of This Work

This study focuses on the refinement of erroneous large intestine 3D surfaces. The contributions of this work can be summarized as follows:

- To the best of our knowledge, we are the first to tackle this problem as a conditional point cloud generation task.
- We construct a dataset consisting of pairs of 3D large intestine shapes. Each pair includes an erroneous shape generated by a deep learning model and the corresponding ground truth.
- Drawing inspiration from LION and PDR, we introduce a novel method for point cloud refinement. Our approach involves training conditional point-diffusion models within a hierarchical latent space.
- Different from PDR which relies on a second network (RFNet), we propose a simple point cloud post-processing pipeline that effectively increases the density and smooths out the noisy clouds before performing surface reconstruction.

	TotalSeg	PET/CT	CAP	Total
Train	216	55	134	405
Validation	32	6	23	61
Test	60	14	38	112
Total	308	75	195	578

Table 1: Dataset final shape counts and split.

3. Material and Methods

3.1. Dataset

We created a dataset that includes pairs of erroneous and correct 3D shapes of the large intestine encoded as voxel grids, polygon meshes and point clouds. Our data include one public and two private datasets.

3.1.1. TotalSegmentator Dataset

Wasserthal et al. (2022) provide a public dataset of 1024 CT scans with reference masks for the large intestine (referred to as "colon"). This class includes the colon (transverse, ascending, descending, and sigmoid), the cecum and the rectum. The dataset contains various CT types, some of which lack the large intestine entirely (e.g., head CTs) or only contain parts of it. We performed connected component analysis to determine the number of components in each mask. Using organ volume distributions derived from the provided reference masks, we established a set of rules to extract the useful cases:

- Cases with a stomach volume lower than 50ml and a urinary bladder volume of 0ml are eliminated, thus ensuring the upper and lower bounds of the intestine are present.
- Outlier cases with large intestine volume lower than 400ml or higher than 1500ml are eliminated.
- Binary closing operations were performed to close small disconnections in the remaining cases that have more than two components (Background+large intestine).
- From the remaining cases, only those with two connected components are kept.

The number of remaining cases after the selection process is 308.

3.1.2. Duke PET/CT Dataset

The set comprises 112 CT volumes from Duke Hospital patients' Positron Emission Tomography and Computed Tomography (PET/CT) scans. We used TotalSegmentator's pre-trained model to obtain segmentation masks for the large intestine. Further analysis for extracting the successful cases involved removing components smaller than 500 voxels and applying binary closing to address minor disconnections. The remaining cases with more than two components were excluded, resulting in 75 remaining cases.

3.1.3. Duke CAP Dataset

The set includes 269 Chest Abdomen Pelvis (CAP) CT scans from Duke Hospital patients. The samples underwent the same pipeline as the PET/CT cases. Additionally, a physician refined 34 cases with incorrect masks resulting in a total of 195 cases.

3.1.4. Partial Shape Synthesis

To train a model that refines erroneous organ shapes, we need a dataset of example failure cases that will be used as model inputs together with the corresponding correct shapes which will be used as the target outputs. Having the previously selected correct shapes, we needed to generate synthetic erroneous masks that mimic the behavior of TotalSegmentator's failures. For this purpose, we built a weak 3D full-resolution U-Net model by training nn-Unet for 30 epochs using 30 images randomly selected from our TotalSegmentator subset. After computing dice scores between the references and the model's outputs, a visual inspection was performed on cases with dice scores below 0.2. It was noted that certain instances had remarkably deficient masks, unlikely to be produced by a properly trained deep learning model. These masks were substituted with their respective complete counterparts. This serves a twofold purpose: enhancing the faithfulness of the synthetic shapes to TotalSegmentator's outputs while teaching the model to accurately reconstruct correct shapes when provided as conditions.

3.1.5. Point Cloud Extraction

After obtaining the pairs of erroneous and reference masks, we run the marching cubes algorithm (Lorensen and Cline (1987)) to extract the organ surfaces represented as polygon meshes. Later, we use the Poisson disk sampling algorithm (Yuksel (2015)) to sample point clouds of 2048 points from the surfaces of the meshes. This algorithm extracts points such that each point has approximately the same distance to the neighboring points. Finally, we normalize the dataset globally to [-1, 1] using the mean and standard deviation calculated over all shapes in the training set.

We split the data into approximately 70% for training, 10% for validation, and 20% for testing. The final numbers of shapes in each set after the split are summarized in table 1.

3.2. Problem Statement

Inspired by Lyu et al. (2021), we formulate our problem as a conditional 3D shape generation task. A 3D point cloud sampled from the surface of a segmentation mask is represented by N points with xyz-coordinates in the 3D space: $x \in \mathbb{R}^{N\times 3}$, We assume the dataset $\mathcal{D}$ is composed of M data pairs $\{(x^i, c^i)|1 \leq i \leq M\}$ where x^i is the i^{th} reference point cloud and c^i is the corresponding erroneous point cloud generated by the weak U-Net model. The goal is to create a conditional model that generates a complete shape y^i, that represents an anatomically plausible shape of a large intestine, using the partial input c^i as a conditioner. Note that the generated shape y^i is as close as possible but does not necessarily match the reference shape x^i. Moreover, due to the stochastic nature of the generation process, the model can produce diverse shapes for different queries with the same conditioner.

3.3. Conditional Generation Network

As a first step, we trained the Conditional Generative Network (CGNet) of PDR on our dataset which will serve as our baseline. Assuming $p_{intestine}$ is the distribution of the complete large intestine shapes x^i and p_{latent} is the latent distribution representing a standard Gaussian in $\mathbb{R}^{N\times 3}$, CGNet is designed as a DDPM that consists of two processes:

- **The forward diffusion process**: a Markov chain that adds noise gradually to the clean data distribution $p_{intestine}$ using Gaussian kernels of fixed variances in T time steps. Variances are fixed such that at the final step T, the shapes belong to the standard Gaussian distribution p_{latent}. This process does not depend on the conditioner.

- **The reverse diffusion process**: a Markov chain implemented as a neural network that predicts and eliminates the noise added during the forward process. This process is conditioned on the erroneous shape c. It starts with a sample x_T from p_{latent} and gradually denoises it to obtain a clean shape x_0 from $p_{intestine}$.

A detailed mathematical formulation of the conditional diffusion procedure is given in Appendix A.

CGNet employs two parallel encoder-decoder networks, namely a condition feature extraction subnet and a denoise subnet. These networks have a shared architecture based on PointNet++. The condition feature extraction subnet extracts multi-level features from the partial input c, while the denoise subnet estimates the noise introduced to the input x_t at time step t. In addition, a two-stage PointNet is used to extract vector-structured global features from the conditioner c. The time step t is embedded into a feature vector using the Transformer's positional embedding (Vaswani et al. (2017)). The global features and the encoded time step are injected into every level of the denoise subnet through shared MLPs. The features extracted by the condition feature extraction subnet are inserted into the parallel level of the denoise subnet via feature transfer modules. The architecture uses attention modules to aggregate the features from the neighboring points to the centers of the clusters in the PointNet++ layers. The variances β_t used during the diffusion process are defined using a linear scheduler.

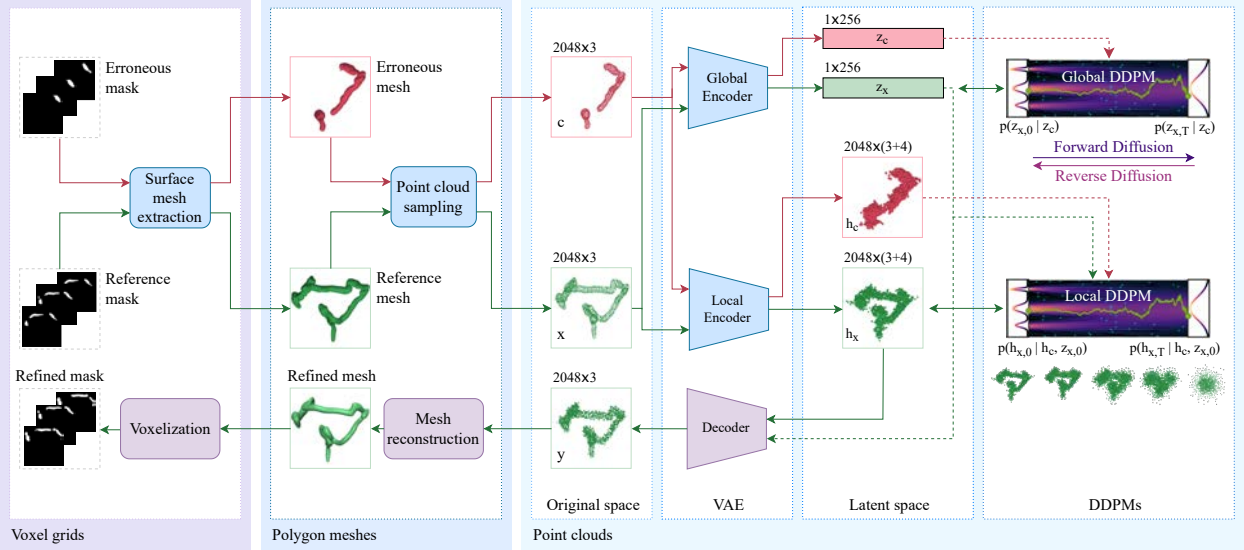


Figure 5: Latent conditional point cloud refinement framework: surface meshes are extracted from the reference and erroneous organ masks using the marching cubes algorithm and point clouds are sampled from both surfaces using Poisson disk sampling. The shapes are encoded into a global and a local representation via two VAE encoders. Two DDPMs are trained to model the distributions of global and local representations of the complete shapes conditioned on the partial shapes' latent representations. The clean shape is reconstructed to the original space via the VAE's decoder. After post-processing, Point-E's deep implicit model is used to reconstruct a polygon mesh that can be voxelized into a 3D mask.

To make the model work on our dataset we introduced some modifications to the default hyperparameters provided by PDR's authors. We used a time embedding of size 128. The depths of all MLPs in the decoders were increased to 3 and the number of neighbors K used for clustering in all PointNet++ layers was set to 10. The number of features of the conditioner (other than the xyz-coordinates) was set to 0. Since we are only training on a single class, the class conditioning mechanism was removed. We trained the model for 8000 epochs with a learning rate of 5×10^{-4} and a batch size of 16.

3.4. Latent Conditional Point-Diffusion Network

To improve the refinement performance, we propose to shift the conditional generation process to a latent space. The inspiration behind this lies in the high complexity and variability of the large intestine's shapes which make it challenging for the DDPM to model their distribution accurately in the original space. We propose to train a VAE that encodes both partial and complete shapes into a unified smoother latent space consisting of a vector-valued global representation and a point cloud-structured local representation. Subsequently, two DDPMs are trained to model the distributions of complete shape latents conditioned on the representations of partial shapes. This enables the disentanglement of high-level features related to the overall appearance of the organ from the low-level features expressing the fine details, which makes it easier for the DDPMs to model the underlying distributions. Finally, the VAE's decoder combines both representations to reconstruct the complete shape in the original space. The overall design of our complete framework is illustrated in Fig. 5.

3.4.1. Hierarchical Latent Shape Encoding

For the latent shape encoding, we adapt the approach proposed by Zeng et al. (2022). Taking a pair of shapes (x, c) from the training set $\mathcal{D}$, the VAE's first encoder learns a global latent representation $(z_x, z_c) \in (\mathbb{R}^{D_z})^2$ where D_z is the size of the latent vector. The second encoder learns a local latent representation $(h_x, h_c) \in (\mathbb{R}^{N\times(3+D_h)})^2$. In other words, h_x and h_c are latent point clouds of N points, each carrying its xyz-coordinates and D_h additional features. By structuring the local latents as point clouds we benefit from the advantages of using this representation in training DDPMs as mentioned in section 2.2.

The VAE consists of two encoders and a decoder all based on the PVCNN architecture. The global encoder takes a 3D point cloud $s \in \{x, c\}$ as input and encodes it into a global latent vector z_s of size D_z. The local encoder takes the point cloud s as input and its global representation z_s as a condition and generates the corresponding local representation h_s consisting of N latent points in $\mathbb{R}^3$ with D_h features. The VAE's decoder takes h_s as input and z_s as a condition and reconstructs the shape s back to the original space.

The VAE is trained by maximizing a modified version of the variational lower bound on the data log-likelihood (ELBO). The loss function combines the data log-likelihood (controlling the shape reconstruction quality) with the Kullback-Leibler (KL) regulariza-

tion (controlling the priors of the latents) with respect to both h_s and z_s. The KL terms of the loss are weighted by two hyperparameters λ_z and λ_h. The VAE is initialized as an identity mapper between the original space, the latent space, and the output to avoid the divergence of the reconstruction loss at the early epochs of training by scaling the variances of the encoders towards 0 and weighting the skip connections accordingly. During the training, the KL weights are annealed linearly until a maximum weight is reached. By increasing the KL weights, the priors of the latents $p(z_s)$ and $p(h_s)$ converge towards a standard Gaussian, making the latent space smoother and more regular, at the cost of increasing reconstruction error.

We fine-tuned the default hyperparameters provided by Zeng et al. (2022) according to our use case. We set D_z to 256 and D_h to 4. The maximum values of both λ_z and λ_h were set to 0.4. To initialize the VAE weights, the variance offset parameter was set to 12 and the skip connections' weight was set to 0.02. The class conditioning mechanism was removed since we are training on a single class. We trained the model using Adam optimizer with a batch size of 32 and a learning rate of 10^{-3} for 8000 epochs while saving the weights every 2000 epochs. Note that the VAE is trained using both partial and complete shapes since both shapes need to be encoded to the latent space for training the DDPMs.

3.4.2. Latent Conditional Point Generation

While freezing the weights of the VAE, we trained two conditional DDPMs in the hierarchical latent space. A first DDPM with parameters ξ was trained on the global latent encodings z_x conditioned on z_c. A second DDPM with parameters ϕ was trained on the local latent encodings h_x conditioned on both h_c and z_x. Similarly to section 3.3, the models are trained by minimizing the difference between the actual noise added to the reference latent encodings and the noise predicted by the DDPMs. The loss functions of the first DDPM can be written as:

$$\mathcal{L}(\xi) = \mathbb{E}_{i\sim\mathcal{U}([M]),t\sim\mathcal{U}([T]),\epsilon\sim\mathcal{N}(0,I)} \left\|\epsilon - \epsilon_\xi(z^i_{x,t}, t, z^i_c)\right\|^2$$

where $\mathcal{U}([M])$ represents the uniform distribution over {1, 2,...,M}, $\mathcal{U}([T])$ represents the uniform distribution over {1, 2,...,T}, $z^i_{x,t}$ is the diffused global latent representation of the shape x^i after t diffusion steps, z^i_c is the global representation of the corresponding conditioner c^i, ϵ is the actual noise and ϵ_ξ is the noise predicted by the model.
Similarly, the loss function of the second DDPM is defined as:

$$\mathcal{L}(\phi) = \mathbb{E}_{i\sim\mathcal{U}([M]),t\sim\mathcal{U}([T]),\epsilon\sim\mathcal{N}(0,I)} \left\|\epsilon - \epsilon_\phi(h^i_{x,t}, t, h^i_c, z^i_{x,0})\right\|^2$$

where $h^i_{x,t}$ is the diffused local latent representation of the shape x^i after t diffusion steps, h^i_c is the local representation of the corresponding conditioner c^i, $z^i_{x,0}$ is the clean global representation of x^i, ϵ is the actual noise and ϵ_ϕ is the noise predicted by the model. The fixed diffusion variances are defined using a linear scheduler for both models. Note that during this process, the latent encodings of the conditioner (z_c, h_c) are only used to extract features that are embedded into the diffusion models to guide the denoising process and they are not diffused.

The global DDPM is implemented as a ResNet with 8 squeeze-and-excitation residual blocks. The network takes a diffused global latent vector $z_{x,t}$ as input and the time step t and the erroneous shape's latent z_c as conditions and outputs the predicted noise. The local DDPM uses the same architecture as the CGNet from section 3.3 with two parallel subnets for condition feature extraction and denoising. The additional PointNet used for extracting global features in the baseline model is removed since we rely on the global representation of the complete shape generated by the global DDPM. This network takes the noisy local representation of a reference shape $h_{x,t}$ as input, it is conditioned on the corresponding local latent of the conditioner h_c, the time step t, and the clean global latent $z_{x,0}$ and it outputs the predicted noise. In both networks, the diffusion step t is encoded using the same positional embedding mechanism used in the baseline.

At inference, the erroneous shape c is encoded into its latent representations (z_c, h_c), and two noisy inputs $z_{x,T}$ and $h_{x,T}$ are sampled from a normal Gaussian distribution. First, the reverse diffusion process is performed on $z_{x,T}$ using the global DDPM to obtain a clean global representation $z_{x,0}$. Later, the local DDPM is used to run the reverse diffusion process on $h_{x,T}$ to obtain a clean local representation $h_{x,0}$ using both h_c and the generated $z_{x,0}$ as conditions. Finally, the resulting representations are decoded back to the original space via the VAE's decoder. This inference process is depicted in Fig. 6.

We used a 256-dimensional time embedding and 1000 diffusion steps in both models. We set the dropout of the ResNet layers to 0.2. Different from the baseline CGNet, the ReLU activation function is replaced by Swish (Ramachandran et al. (2017)). Based on the feature size of our local latent representations, the number of the partial input features is set to 4 and the feature dimension of the output is set to 7. The DDPMs are trained in parallel using Adam optimizer with a batch size of 10 and a learning rate of 2×10^{-4} for 16000 epochs. The remaining training hyperparameters are set to the default values provided by Zeng et al. (2022).

3.5. Point Cloud Post-Processing

The output of the generative model tends to be sparse and noisy which effects negatively the performance of the surface reconstruction. To address this issue we propose a simple point cloud post-processing pipeline consisting of the following steps:

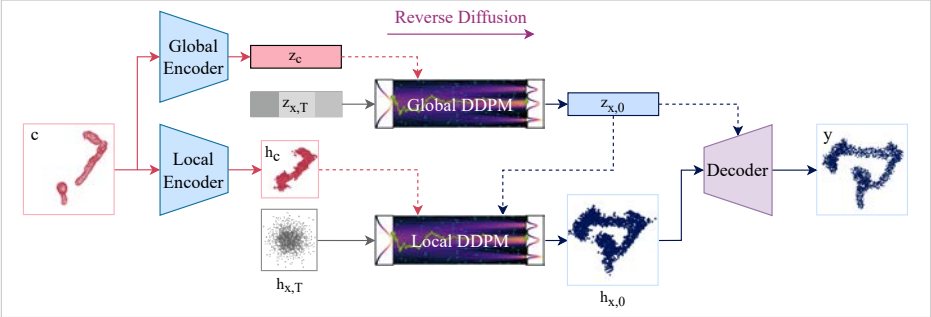


Figure 6: Inference pipeline: The erroneous shape c is encoded into latent representations (z_c, h_c), and noisy inputs $z_{x,T}$ and $h_{x,T}$ are sampled. Reverse diffusion processes are applied using the global and local representations of c as conditions to obtain clean representations $z_{x,0}$ and $h_{x,0}$. These representations are then decoded back to the original space using the VAE's decoder.

- The point cloud is first renormalized back to the original scale.

- The cloud is smoothed using a variant of the Moving Least Square algorithm with a smoothing factor of 0.2.

- The cloud is densified by adding new points under the assumption that all points in a local neighborhood are within a specified distance of each other. If any neighbor exceeds the target distance, the connecting edge is divided and a new point is inserted at the midpoint. We use a target distance of 10 mm and a neighborhood size of 10 points. We apply this process for one iteration.

- Outlier removal is applied such that all points having less than 5 neighbors within a radius of 15 mm are removed.

Note that after this process, the resulting point clouds will have different numbers of points depending on the quality of the raw generated output.

3.6. Surface Reconstruction

Since the goal is to generate organs for computational phantoms that are usually represented as polygon meshes, we extend our method with a deep implicit model for point-cloud-to-mesh reconstruction (pc2mesh) proposed within the Point-E framework (Nichol et al. (2022)). This model uses an encoder-decoder Transformer architecture and predicts SDFs based on the input point clouds. The mesh is obtained by applying the marching cubes algorithm on the generated SDF map. The model is trained with 2.4M meshes. In our experiments, we used the provided pre-trained weights and a grid size of $128 \times 128 \times 128$ with a batch size of 1024 points. Point clouds were normalized to [-0.5, 0.5] before being fed to the model. The reconstructed meshes can then be binarized to obtain a voxel grid.

4. Results

4.1. Evaluation Metrics

4.1.1. Chamfer Distance

Chamfer Distance (CD) is the most frequently used metric for evaluating 3D shape completion performance. CD tries to find the minimum distance between two sets of points. For a reference shape X and a generated shape Y, the Chamfer distance between the two shapes is defined as follows:

$$\mathrm{CD}(X,Y) = \frac{1}{|X|}\sum_{x\in X}\min_{y\in Y}\|x-y\|_2^2 + \frac{1}{|Y|}\sum_{y\in Y}\min_{x\in X}\|y-x\|_2^2$$

CD provides information about the overall similarity and alignment between the two point clouds as well as information about local similarity since it relies on pairwise distances between individual points.

4.1.2. Hausdorff Distance

The Hausdorff distance (HD) provides a measure of the largest dissimilarity between the two point clouds. It measures the maximum distance of a point in one cloud to its nearest point in the other cloud. Formally, the Hausdorff distance between X and Y is defined as:

$$\mathrm{HD}(X,Y) = \max\left\{\sup_{x\in X}\inf_{y\in Y}\|x-y\|, \sup_{y\in Y}\inf_{x\in X}\|y-x\|\right\}$$

where sup represents the supremum and inf the infimum. This metric highlights the worst-case scenario and provides an upper bound on the error between the completed and reference point clouds.

4.1.3. Earth Mover Distance

The Earth Mover distance (EMD), also known as the Wasserstein distance, is a metric used to quantify the dissimilarity between two point clouds. It measures the minimum cost required to transform one point cloud into another, where each point is treated as a mass that needs to be moved. Given two point clouds, X and Y, with N points each, the EMD can be computed using the following formula:

$$\mathrm{EMD(X,Y)} = \min_{\gamma: Y\longleftrightarrow X}\sum_{y\in Y}\|y-\gamma(y)\|_2$$

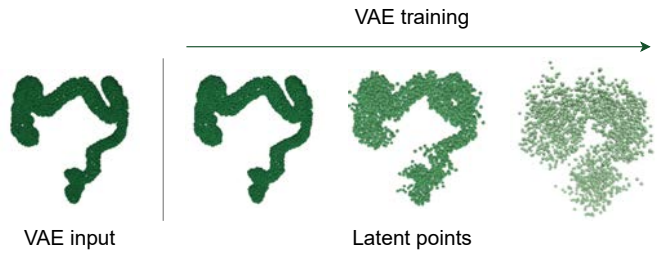


Figure 7: Latent points evolution during the VAE training.

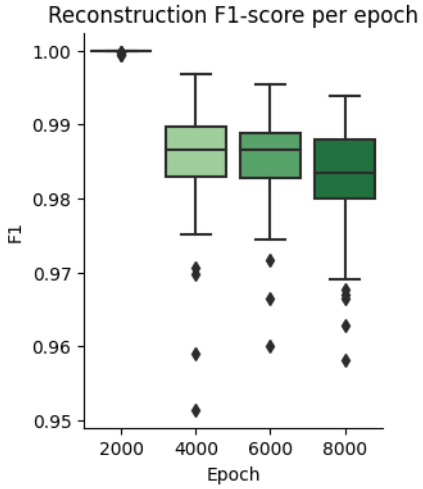


Figure 8: Reconstruction quality evaluation during the VAE training in terms of F1-Score.

where γ is a bijection between the point clouds X and Y. EMD accounts for both the global geometry and the mass distribution of the points.

CD, HD and EMD collectively enable the assessment of completeness, geometric accuracy, and distribution fidelity between the generated point clouds and the reference shapes.

4.1.4. F1-Score

To evaluate the point cloud reconstruction performance of the VAE we used the F1-score, as proposed by Tatarchenko et al. (2019). It combines precision and recall, representing the reconstruction's accuracy and completeness respectively. The F1-score can be adjusted using a distance threshold (d) to control its strictness. For a reference shape X and a reconstructed shape Y, it is defined as:

$$F1_d(X,Y) = 2 \times \frac{P_d(X,Y)R_d(X,Y)}{P_d(X,Y) + R_d(X,Y)}$$

where P is the precision:

$$P_d(X,Y) = \frac{1}{|Y|}\sum_{y \in Y}[\min_{x \in X} \|x - y\| < d]$$

and R is the recall:

$$P_d(X,Y) = \frac{1}{|X|}\sum_{x \in X}[\min_{y \in Y} \|x - y\| < d]$$

4.2. Latent Shape Encoding and Reconstruction

As explained in section 3.4.1, there is a trade-off between the regularization of the latent space and the reconstruction performance of the hierarchical VAE. Fig. 7 illustrates an example of how the latent points evolve during the VAE training. We analyzed the reconstruction performance on our test set for different checkpoints of the VAE to choose the weights to be used for training the DDPMs. The boxplots in Fig 8 summarize the reconstruction performance of the VAE per epoch in terms of F1-score. A threshold of d=5mm was used to compute the metric[1].

As expected, during the early epochs of the training the latent points have a similar shape to the input resulting in a good reconstruction performance reflected in the high F1-score, but the complex distribution of these latent points is not convenient for training the DDPMs. As the training continues, the latent points start to exhibit smoother and more regularized distributions that approach a standard Gaussian at the latest epochs, which is preferred for training the DDPMs. But the reconstruction performance decreases as indicated by the lower F1-score means and higher dispersion. Based on the results shown above and a qualitative evaluation of the reconstruction performance and the smoothness of the latent points, the weights of the VAE at epoch 6000 were used for training the DDPMs in the next stage. A more detailed study on the impact of the selected weights is presented in section 4.7.2.

4.3. Point Cloud Refinement

After running the inference pipeline using both CGNet and our proposed latent conditional point diffusion model on our synthetic test set, we summarize the results of point cloud refinement obtained in terms of CD, HD, and EMD in table 2. We also report the metrics obtained initially between the complete and the erroneous shapes (used as model conditions) for reference. Fig 9 shows examples of refined large intestine point clouds using both models.

[1]Note that the F1−score interval was cropped to [0.95,1] to better visualize the difference between the plots.

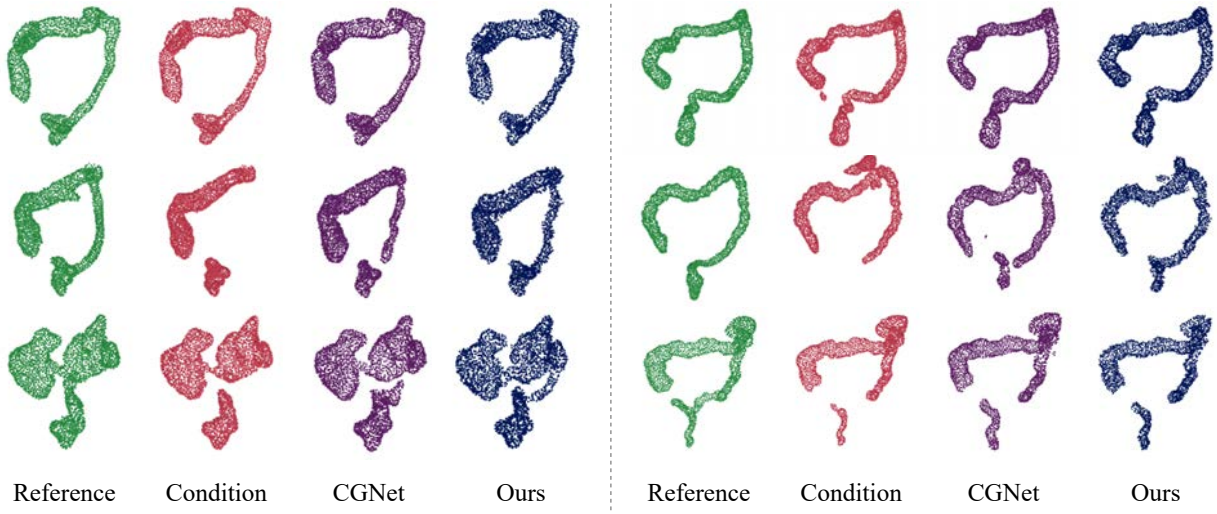


Figure 9: Examples of the shape refinement results

Model	CD ↓	HD ↓	EMD ↓
Init	1396 ± 2615	84.01 ± 59	9430 ± 7150
CGNet	449 ± 853	57.18 ± 32	**9294** ± **4940**
Ours	**388** ± **681**	**56.44** ± **31**	9666 ± 4855

Table 2: Shape refinement results on the test set.

We can see that both the CGNet and our proposed latent conditional point-diffusion model exhibit significant improvements in the surface reconstruction performance compared to the starting point with our model outperforming CGNet with a 13.95% improvement in CD and 1.29% in HD on average, indicating higher performance in handing local errors and a better alignment with the reference shapes. In contrast, the CGNet demonstrates superior performance compared to our pipeline in terms of EMD. This indicates that the shapes generated by our model exhibit lower visual quality and less uniform density. Furthermore, the latent model demonstrates reduced dispersion, as indicated by the lower standard deviation for all metrics indicating higher stability. Qualitatively, our method accurately captures the overall distribution of large intestine shapes, generating anatomically acceptable results. It effectively completes missing parts and eliminates false positives in certain scenarios. In addition, when given a complete shape as a condition, we can see that both models accurately reconstruct the complete point cloud. However, shapes generated by the latent model tend to be noisier and sparser compared to CGNet as reflected in the higher EMD values. Additionally, the model struggles to remove adjacent or attached false positives but often combines them with the correct segments generating a more plausible connected representation of a large intestine. Occasionally, the model fails to connect organ segments.

4.4. Surface Reconstruction

The generated point clouds of both CGNet and our model were post-processed using the pipeline proposed in section 3.5. The point clouds were then given as input to Point-E's point-cloud-to-mesh model to generate the corresponding polygon meshes. To evaluate the performance, we sampled 50000 points from the mesh surfaces and computed CD, HD, and EMD[2] between the generated and the reference meshes. The results are summarized in table 3. Examples of the post-processed point clouds and the corresponding reconstructed meshes are shown in Fig. 10.

Model	CD ↓	HD ↓	EMD ↓
Init	1400 ± 2620	84.22 ± 58	9767 ± 6995
CGNet	463 ± 871	57.13 ± 32	9273 ± 4950
Ours	**409** ± **716**	**56.43** ± **31**	**9096** ± **4836**

Table 3: Shape refinement results after surface reconstruction.

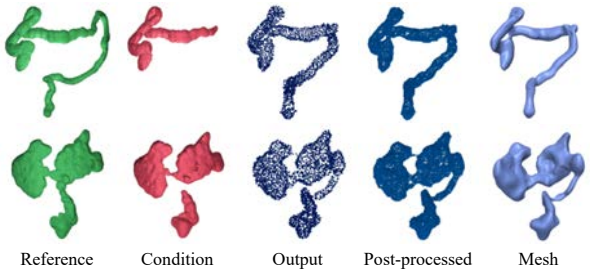


Figure 10: Example results of post-processed point clouds and the corresponding reconstructed meshes.

[2]2048 Points were used to compute EMD due to GPU memory limits.

Representation	Model	CD ↓	HD ↓	EMD ↓
Point clouds	TotalSeg	579 ± 938	93.62 ± 137	**6545 ± 3697**
	TotalSeg+CGNet	318 ± 588	**75.03 ± 130**	9011 ± 5071
	TotalSeg+Ours	**220 ± 321**	**73.16 ± 132**	8691 ± 3080
Meshes	TotalSeg	571 ± 929	93.97 ± 137	**6916 ± 3766**
	TotalSeg+CGNet	313 ± 543	74.81 ± 131	8432 ± 4543
	TotalSeg+Ours	**230 ± 346**	**72.46 ± 132**	7893 ± 3144

Table 4: Shape refinement results on TotalSegmentator cases

After applying post-processing and surface reconstruction, our method maintains high performance across all metrics. Notably, our model outperforms CGNet in all metrics, including EMD, indicating the effectiveness of our post-processing and mesh reconstruction pipeline in reducing the noise and sparsity of the generated point clouds. This highlights the ability of our pipeline in modeling the distribution of the large intestine's shape in terms of global appearance, local details, and density uniformity. Additionally, visual inspection reveals that the post-processed point clouds exhibit smoother and denser characteristics compared to the raw ones. The reconstructed meshes exhibit high quality and preserve fine details. In fact, they are less affected by discretization compared to the reference meshes generated from the binary segmentation masks. However, in case of adhesion between the colon's walls, the generated mesh does not accurately represent this phenomenon. Instead, the adhesive segments are merged into a single, larger tube. It should be noted that this error is inherited from the training set, where applying the marching cubes algorithm to the voxelized mask produces the same defect (refer to the reference mesh of the lower row in Fig. 10).

Additional qualitative results are reported in Appendix B.

4.5. TotalSegmentator Refinement

Following the evaluation of our proposed pipeline on the synthetic test set, we further conducted an experiment to assess its performance on real-world data. Specifically, we examined its ability to handle failure cases generated by the state-of-the-art segmentation model, TotalSegmentator. Along with our synthetic dataset, we acquired a set of 20 additional CT scans (8 CAP and 12 PET/CT) and generated the segmentation masks of the large intestine using TotalSegmentator which will serve as our conditioner input of the model. A physician refined the segmentation masks generating the corresponding reference masks. Our pipeline was applied to this dataset using both CGNet and the latent conditional point diffusion model. Table 4 shows the refinement performance of the raw generated point clouds and the reconstructed meshes. Visual examples of the generated shapes can be found in Appendix C.

The results demonstrate significant improvement of the proposed method in refining the large intestine's surface compared to the segmentation results of TotalSegmentator. On average, we achieved a 62.32% improvement in CD and a 2.15% improvement in HD, outperforming CGNet in both metrics. However, it is worth noting that both our proposed method and CGNet exhibit poor EMD results compared to TotalSegmentator, although our model still outperforms CGNet in this metric. The qualitative evaluation confirms that the observations made on the synthetic set (accurate distribution modeling, noise elimination in certain cases, and failure scenarios) are consistent with real-world cases.

4.6. Computational Time Analysis

The VAE's training time was 38 hours distributed on 2 NVIDIA RTX A6000 GPUs whereas the training of the DDPMs lasted for 63 hours distributed on 4 NVIDIA RTX A6000 GPUs. We computed an estimation of the average inference time of each component of our refinement pipeline on a set of 20 samples. The results are shown in table 5.

Process	Computational time (s)
Marching cubes	4.39 ± 3.5
Poisson disk sampling	0.09 ± 0.0
VAE+DDPM inference	96.83 ± 1.8
Post-processing	0.44 ± 0.1
Point-E pc2mesh	9.80 ± 2.2
Complete pipeline	111.555 ± 4.5

Table 5: Average computational time required for each process.

On average, the entire pipeline can be executed in less than 2 minutes for a single case. It is worth noting that the latent DDPM sampling process consumed the most time, as expected, due to the execution of 1000 reverse diffusion steps for each of the two diffusion models. It is important to highlight that the process was individually applied to each sample. Processing the clouds in batches during the execution of the reverse diffusion processes and the mesh reconstruction could significantly decrease the average time for a larger set of samples.

Model	CD ↓	HD ↓	EMD ↓
A	**431 ±715**	**55.45 ±30**	11364 ±4731
B	455 ±785	57.44 ±33	**9674 ±5321**

Table 6: Shape refinement results for the uniform sampling-based model (A) and Poisson disk sampling-based model (B).

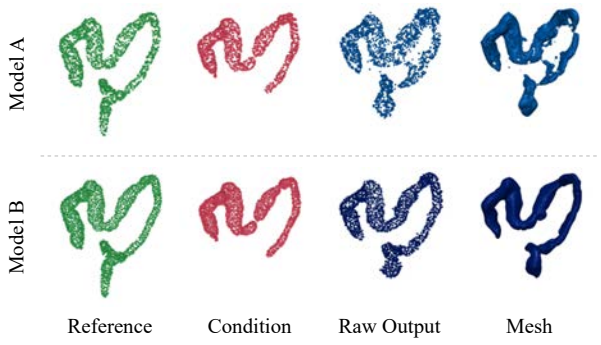


Figure 11: An example illustrating the impact of the point sampling strategy on the generated point clouds.

Model	CD ↓	HD ↓	EMD ↓
C	422 ± 847	55.99 ±32	10231 ±4913
D	**388 ±681**	56.44 ±31	**9666 ±4855**
E	411 ±821	**54.25 ±32**	9972±4975

Table 7: Experimental results on latent space smoothness impact on the generated organ point clouds.

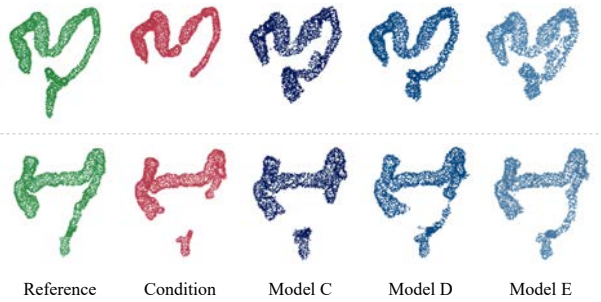


Figure 12: Examples illustrating the impact of latent space smoothness on the generated organ point clouds.

4.7. Ablation Study

4.7.1. Point Sampling Strategy

We studied the impact of using Poisson disk sampling to extract the points from the mesh surfaces compared to uniformly sampling points based on triangle area. We trained two models A and B where in A we sample the points uniformly and in B we use Poisson disk sampling. The VAEs of both models were trained for 6000 epochs whereas the DDPMs were trained for 10000 epochs. The results on the test set are calculated after reconstructing the mesh surfaces and are summarized in table 6. An example of the generated results is shown in Fig. 11.

The results show that the uniform sampling-based model slightly outperforms the Poisson disk-based model in terms of CD and HD whereas the latter achieves better EMD. As illustrated in Fig 11, The point clouds sampled uniformly exhibit heterogeneous density across the surface whereas the ones sampled using the Poisson disk sampling are more uniform. The results of the model trained with uniformly sampled clouds suffer more dispersion in the generated point clouds compared to the Poisson disk-based model, which impacts negatively the quality of the generated meshes explaining the poor performance in terms of EMD.

4.7.2. Latent Space Smoothness Impact

To evaluate the impact of the selected VAE's weights on the generation process, we trained three latent DDPMs C, D, and E using the VAE weights saved at epochs 4000, 6000, and 8000 respectively. The generation metrics obtained on the raw point clouds are reported in table 7. Example shapes generated using the different models are illustrated in Fig 12.

The results show that the model trained with the VAE weights of epoch 8000 outperformed the other models in terms of both CD and EMD whereas the model trained with the VAE weights at epoch 6000 achieves the lowest HD among the three models. The qualitative evaluation revealed that by extending the VAE's training time (therefore smoothing the latent space), the DDPMs capture better the distribution of the shapes and the organs generated are more compact and connected. However, the reconstruction quality decreases, and the generated shapes tend to be noisier. Overall, using the VAE weights at epoch 6000 balances the trade-off between the latent space smoothness and the reconstruction performance.

4.7.3. Post-processing

To assess the impact of the proposed post-processing pipeline, we generated meshes using both the raw and the post-processed point clouds and we report the results we obtained in table 8. This study was conducted on the results of the model trained with VAE weights at epoch 8000 which exhibit more noise, sparsity, and outliers. Examples of the generated meshes from the raw and post-processed clouds are illustrated in Fig 13.

While the performance metrics do not exhibit a substantial difference, the qualitative evaluation highlights the benefits of post-processing. The meshes generated after post-processing are smoother and suffer fewer holes and noise along their surfaces, which is more compatible with the anatomy of the large intestine's walls primarily consisting of soft tissues.

Post-process	CD ↓	HD ↓	EMD ↓
Before	**427 ±837**	54.07 ±32	**9344 ±4914**
After	431 ±849	**54.05 ±32**	**9253 ±5032**

Table 8: Mesh reconstruction results before and after post-processing.

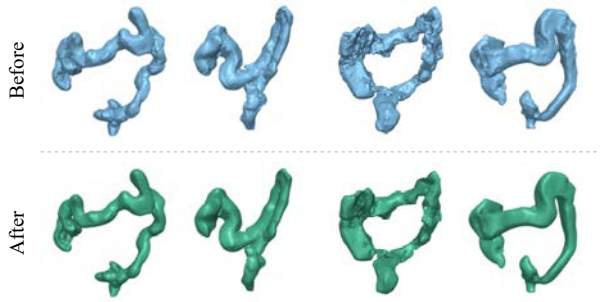


Figure 13: Examples illustrating the impact of post-processing on the reconstructed meshes.

4.7.4. Data Augmentation

To address the problem of small dataset size, we implemented a simple data augmentation pipeline based on rigid 3D transformations. We use a scaling factor of up to 10% of the original size, a rotation around the z-axis of up to 10°, and a translation of up to 0.1 units in the normalized space. We apply this pipeline to train both the VAE and DDPMs of a new model (Ours-Aug). The quantitative results computed from the generated point clouds are represented in table 9. Figure 14 illustrates examples generated using this model.

Compared to the models trained without data augmentation, we can see that this model performs better in addressing some of the issues encountered in the previous models, such as reducing false positives and generating shapes that closely match the corresponding reference shapes, resulting in minimized HD. However, the generated shapes exhibit more dispersion and noise and can generate errors that were not commonly seen in the previous models, such as the presence of links between the cecum and the rectosigmoid junction, leading to higher EMD and CD values.

Model	CD ↓	HD ↓	EMD ↓
Ours-Aug	449 ± 814	54.62 ± 32	11019 ± 4874

Table 9: Shape refinement results of the model trained with data augmentation

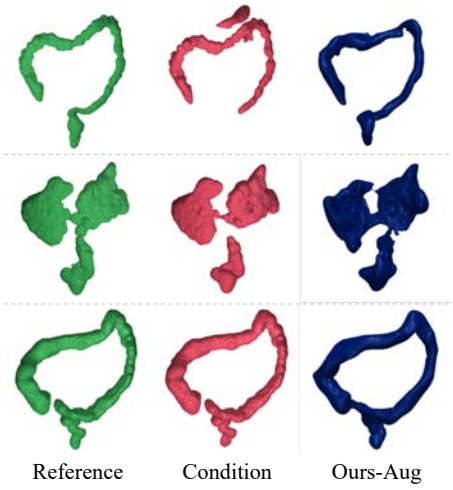


Figure 14: Example shapes generated by the model trained with data augmentation.

5. Discussion

In this study, we investigated the application of geometric deep learning techniques and denoising diffusion models to refine erroneous segmentation masks of the large intestine. These masks were generated by a volumetric segmentation model and exhibit multiple issues such as missing parts or noise. We approached the problem as a conditional point cloud generation task and proposed a latent conditional point diffusion model for point cloud refinement. Our pipeline involves sampling point clouds from the organ's surfaces and encoding the shapes into a smoother latent space consisting of a global and a local representation using a hierarchical VAE's encoder. Two DDPMs are then trained in this latent space to perform point cloud refinement. The generated latent point clouds are then decoded back to the original space using the VAE's decoder before being post-processed. Finally, a surface mesh is reconstructed using a modern deep implicit model.

By comparing our proposed method to CGNet, we observed that training the DDPMs in a hierarchical latent space yields significant advantages in modeling shape distributions, particularly in capturing fine details. Our method demonstrates superior local performance by effectively handling the false positives, while CGNet tends to produce anatomically inaccurate details, such as including small branches in some sections of the colon. Furthermore, our method demonstrated fine-grained precision in shape completion and consistent proximity of the completed points to the reference. Our conclusions were reinforced by the study on the impact of latent space smoothness, which showed that models trained in smoother spaces better preserved the organ's anatomy. However, it's important to note that CGNet generated denser and more uniform point clouds compared to the latent model which tended to generate noisier and sparser point clouds. This phenomenon is attributed to the encoding and decoding process between the original and latent spaces.

Our proposed post-processing and mesh reconstruction methods effectively resolved this issue, producing high-quality meshes with detailed structures and minimal noise. Further fine-tuning of the post-processing

parameters could enhance the shape quality. Additionally, using Poisson disk sampling for point extraction resulted in outputs with more uniform point density across the organ's surface. This finding suggests that enforcing uniformity in the input enables the model to learn evenly distributed points, capturing the characteristic of even distances between points as a feature of the desired output.

Unlike Balsiger et al. (2019) who suggests using a point classification network to eliminate background points in the segmentation result, our method utilizes generative deep learning. This enabled our model to generate new parts and complete partial and disconnected shapes obtained from the segmentation model. Friedrich et al. (2023), on the other hand, adopted PVD's architecture (Zhou et al. (2021)) for generating skull implants. Their method assumes that the conditioner is entirely contained within the target shape and maintains it fixed throughout the diffusion process. However, in our case, the conditioner may contain noise and fragments of other organs, rendering the use of such techniques impractical since the false positive points will still be present in the output. To overcome this issue, we employed the CGNet architecture introduced by Lyu et al. (2021). Unlike the model used in PVD, CGNet does not fix the partial shape but rather extracts its features to guide the generation process. By adopting this strategy, our model successfully eliminated noisy components from the masks in several cases.

We applied our pipeline to outputs generated by TotalSegmentator, a well-known model that was trained on one of the largest imaging datasets. Compared to the corresponding masks refined by a physician, we noticed a significant improvement in the surface representation performance. Additionally, the qualitative evaluation showed consistency with the results obtained on our test set indicating the faithfulness of our synthetic dataset to the actual outputs of the model and the generalizability of our model to new cases. On the other hand, the model resulted in higher EMD values compared to TotalSegmentator's output. Besides model failures, this observation can be attributed to several factors. For instance, the cases included in this set suffered mostly from local problems such as disconnections or small parts included from the other organs and EMD does not quantify well local dissimilarity. Moreover, the reference and TotalSegmentator's point clouds were both generated using the same procedure (an application of the marching cubes algorithm followed by Poisson disk sampling) which explains the higher similarity between the density distributions of both sets.

5.1. Limitations

Despite demonstrating promising results, our method for large intestine segmentation refinement still has certain limitations. Firstly, the model encounters difficulties in effectively eliminating false positives that are closely adjacent or attached to the actual segments of the organ. This challenge arises because PointNet's layers classify these points as neighbors of nearby true positives, leading to their inclusion in the clusters and the contribution of their features to the generation process.

Secondly, the model occasionally struggles to connect segments of the large intestine, particularly when the missing portion lies in the rectosigmoid junction. Besides, the model can produce anatomically inaccurate shapes when the partial input is complex or exhibits multiple curvatures. This phenomenon is attributed to the high variability in the organ's shapes across patients and even within the same patient, due to its dynamic filling process. Expanding the size of the training set may help the model to adequately cover the complex distribution of the organ's shape.

Additionally, the generated shapes do not always align perfectly with the ground truth and may intersect with neighboring organs, preventing the 3D model from fitting in the corresponding computerized phantom. This limitation arises because the current model is only guided by the partial shape provided as a conditioner, lacking crucial contextual information about the surrounding organs.

5.2. Future Work

In future work, addressing the problem of false positives can be achieved through further engineering of the neighborhood definition hyperparameters and attention modules within the CGNet model. These enhancements would enable the model to better distinguish between true positives and false positives, ultimately improving the refinement process.

To address the challenge of limited data availability, one approach is to acquire more scans to increase diversity. Additionally, a better selection of the data augmentation parameters and sampling multiple partial shapes from the same segmentation mask would prove beneficial in providing the model with a more comprehensive understanding of organ shape distribution and potential segmentation model failures.

To avoid intersecting with neighboring organs and improve the accuracy of the generated shapes, extracting landmarks from the neighboring organs and using them as a second condition of the generative models can be explored. This approach would help the model restrict the region of generation and ensure better spatial alignment with the patient's phantom.

Expanding the model to handle multiple organs can be accomplished by including their shapes to the dataset as an additional class. However, the model cannot distinguish between organs if both masks are included in the same data sample, extending the model with point classification heads would make the model label-aware and enable it to refine multiple organs simultaneously.

Currently, we depend on a separate pre-trained model, Point-E, for mesh reconstruction. To enhance

the reconstruction performance, we could fine-tune the model on our data. Another viable approach would be modifying the decoder of the VAE to directly output signed distance functions. This adjustment would simplify the process of mesh reconstruction, avoiding the generation of sparse and noisy point clouds.

By implementing these solutions, we aim to enhance the model's performance in terms of false positive elimination, dataset diversity, refinement of multiple organs, and mesh reconstruction quality.

6. Conclusions

We have presented an end-to-end automatic pipeline for refining 3D shapes of the large intestine, which improves the surface reconstruction of the organ starting from an erroneous segmentation. Our method is based on geometric deep learning and denoising diffusion probabilistic models. We formulate the refinement process as a conditional point cloud generation problem performed in a hierarchical latent space. Through a comprehensive evaluation of the method on both our synthetic test set and real-world cases, our approach has demonstrated promising results both quantitatively and qualitatively. The refined 3D shapes exhibited improved surface reconstruction and enhanced anatomical accuracy compared to the outputs of the segmentation model. This study validates the effectiveness of utilizing geometric DL and DDPMs in enhancing the surface reconstruction of deformable anatomical structures, using the large intestine as an example.

Our method opens up possibilities for further improvement and can be extended to multiple applications that could enhance the quality of computerized phantoms. Future work includes incorporating additional contextual information from neighboring structures to restrict the generation region and expanding the model to perform label-aware refinement of multiple organs.

Acknowledgments

I would like to extend my deepest gratitude to my supervisor, Dr. Joseph Lo, for not only providing me with this incredible opportunity but also for his support, invaluable supervision, and exceptional guidance. I am immensely grateful to Duke CVIT for hosting this project and providing the necessary computational resources for its execution. I would like to thank Dr. Mobina Ghojogh Nejad and Lavsen Dahal for their insightful contributions, as well as Xiaohui Zeng, the author of LION, for his assistance and sharing his expertise. I extend my sincere appreciation to the MaIA master consortium and the European Commission for granting me this life-changing opportunity and generously funding my education. Lastly, I would like to thank my family and friends for their lasting support and patience.

References

Balsiger, F., Soom, Y., Scheidegger, O., Reyes, M., 2019. Learning shape representation on sparse point clouds for volumetric image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22, Springer. pp. 273–281.

Cai, J., Xia, Y., Yang, D., Xu, D., Yang, L., Roth, H., 2019. End-to-end adversarial shape learning for abdomen organ deep segmentation, in: Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10, Springer. pp. 124–132.

Cerrolaza, J.J., Picazo, M.L., Humbert, L., Sato, Y., Rueckert, D., Ballester, M.Á.G., Linguraru, M.G., 2019. Computational anatomy for multi-organ analysis in medical imaging: A review. Medical image analysis 56, 44–67.

Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y., 2022. Transmorph: Transformer for unsupervised medical image registration. Medical image analysis 82, 102615.

Chou, G., Bahat, Y., Heide, F., 2022. Diffusionsdf: Conditional generative modeling of signed distance functions. arXiv preprint arXiv:2211.13757 .

Dahal, L., Wang, Y., Tushar, F.I., Montero, I., Lafata, K.J., Abadi, E., Samei, E., Segars, W.P., Lo, J.Y., 2023. Automatic quality control in computed tomography volumes segmentation using a small set of xcat as reference images, in: Medical Imaging 2023: Physics of Medical Imaging, SPIE. pp. 857–861.

Fan, H., Yang, Y., 2019. Pointrnn: Point recurrent neural network for moving point cloud processing. arXiv preprint arXiv:1910.08287 .

Friedrich, P., Wolleb, J., Bieder, F., Thieringer, F.M., Cattin, P.C., 2023. Point cloud diffusion models for automatic implant generation. arXiv preprint arXiv:2303.08061 .

Hesterman, J.Y., Kost, S.D., Holt, R.W., Dobson, H., Verma, A., Mozley, P.D., 2017. Three-dimensional dosimetry for radiation safety estimates from intrathecal administration. Journal of Nuclear Medicine 58, 1672–1678.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems 33, 6840–6851.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods 18, 203–211.

Kong, F., Wilson, N., Shadden, S., 2021. A deep-learning approach for direct whole-heart mesh reconstruction. Medical Image Analysis 74, 102222. doi:https://doi.org/10.1016/j.media.2021.102222.

Lee, C., Lodwick, D., Hasenauer, D., Williams, J.L., Lee, C., Bolch, W.E., 2007. Hybrid computational phantoms of the male and female newborn patient: Nurbs-based whole-body models. Physics in Medicine & Biology 52, 3309.

Li, Y., Bu, R., Sun, M., Wu, W., Di, X., Chen, B., 2018. Pointcnn: Convolution on x-transformed points. Advances in neural information processing systems 31.

Liu, Y., Fan, B., Xiang, S., Pan, C., 2019a. Relation-shape convolutional neural network for point cloud analysis, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8895–8904.

Liu, Y., Lei, Y., Fu, Y., Wang, T., Tang, X., Jiang, X., Curran, W.J., Liu, T., Patel, P., Yang, X., 2020. Ct-based multi-organ segmentation using a 3d self-attention u-net network for pancreatic radiotherapy. Medical physics 47, 4316–4324.

Liu, Z., Tang, H., Lin, Y., Han, S., 2019b. Point-voxel cnn for efficient 3d deep learning. Advances in Neural Information Processing Systems 32.

Lorensen, W.E., Cline, H.E., 1987. Marching cubes: A high resolution 3d surface construction algorithm. ACM siggraph computer graphics 21, 163–169.

Luo, S., Hu, W., 2021. Diffusion probabilistic models for 3d point cloud generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2837–2845.

Lyu, Z., Kong, Z., Xu, X., Pan, L., Lin, D., 2021. A conditional point diffusion-refinement paradigm for 3d point cloud completion. arXiv preprint arXiv:2112.03530 .

Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., Chen, M., 2022. Point-e: A system for generating 3d point clouds from complex prompts. arXiv preprint arXiv:2212.08751 .

Oda, H., Hayashi, Y., Kitasaka, T., Tamada, Y., Takimoto, A., Hinoki, A., Uchida, H., Suzuki, K., Itoh, H., Oda, M., et al., 2021. Intestinal region reconstruction of ileus cases from 3d ct images based on graphical representation and its visualization, in: Medical Imaging 2021: Computer-Aided Diagnosis, SPIE. pp. 388–395.

Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 652–660.

Raju, A., Miao, S., Jin, D., Lu, L., Huang, J., Harrison, A.P., 2022. Deep implicit statistical shape models for 3d medical image delineation, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 2135–2143.

Ramachandran, P., Zoph, B., Le, Q.V., 2017. Searching for activation functions. arXiv preprint arXiv:1710.05941 .

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer. pp. 234–241.

Segars, W., Bond, J., Frush, J., Hon, S., Eckersley, C., Williams, C.H., Feng, J., Tward, D.J., Ratnanather, J., Miller, M., et al., 2013. Population of anatomically variable 4d xcat adult phantoms for imaging research and optimization. Medical physics 40, 043701.

Segars, W.P., Sturgeon, G., Mendonca, S., Grimes, J., Tsui, B.M., 2010. 4d xcat phantom for multimodality imaging research. Medical physics 37, 4902–4915.

Tatarchenko, M., Richter, S.R., Ranftl, R., Li, Z., Koltun, V., Brox, T., 2019. What do single-view 3d reconstruction networks learn?, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Vahdat, A., Kreis, K., Kautz, J., 2021. Score-based generative modeling in latent space. Advances in Neural Information Processing Systems 34, 11287–11302.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

Wang, C., Cui, Z., Yang, J., Han, M., Carneiro, G., Shen, D., 2022. Bowelnet: Joint semantic-geometric ensemble learning for bowel segmentation from both partially and fully labeled ct images. IEEE Transactions on Medical Imaging .

Wang, M., Guo, N., Hu, G., El Fakhri, G., Zhang, H., Li, Q., 2016. A novel approach to assess the treatment response using gaussian random field in pet. Medical Physics 43, 833–842.

Wasserthal, J., Meyer, M., Breit, H.C., Cyriac, J., Yang, S., Segeroth, M., 2022. Totalsegmentator: robust segmentation of 104 anatomical structures in ct images. arXiv preprint arXiv:2208.05868 .

Weston, A.D., Korfiatis, P., Philbrick, K.A., Conte, G.M., Kostandy, P., Sakinis, T., Zeinoddini, A., Boonrod, A., Moynagh, M., Takahashi, N., et al., 2020. Complete abdomen and pelvis segmentation using u-net variant architecture. Medical physics 47, 5609–5618.

Yang, J., Wickramasinghe, U., Ni, B., Fua, P., 2022. Implicitatlas: learning deformable shape templates in medical imaging, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15861–15871.

Yuksel, C., 2015. Sample elimination for generating poisson disk sample sets, in: Computer Graphics Forum, Wiley Online Library. pp. 25–32.

Zeng, X., Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S., Kreis, K., 2022. Lion: Latent point diffusion models for 3d shape generation, in: Advances in Neural Information Processing Systems.

Zhou, L., Du, Y., Wu, J., 2021. 3d shape generation and completion through point-voxel diffusion, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5826–5835.

Appendix A. Denoising Diffusion Probabilistic models: Mathematical formulation

Assuming a dataset $\mathcal{D}$ of M shape pairs (x^i, c^i), DDPMs are designed in two steps. The forward diffusion process is implemented as a Markov chain with a fixed variance schedule which adds noise to the input iteratively in a total of T time steps. In the following formulation the superscript i is omitted for simplicity:

$$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{A.1}$$

where $q(x_t|x_{t-1})$ is a Gaussian kernel defined as:

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t;\ \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{A.2}$$

The variances β_t are selected such that the chain converges to a normal Gaussian distribution at step T, $q(x_T) \simeq \mathcal{N}(x_T; 0, I)$.

From eq. A.1 we can see that this process is recursive and the value at time step t depends on the previous steps 0 through $t-1$. To avoid calculating all the intermediate steps, a closed-form expression is given for obtaining the noisy shape x_t at any step t depending on only the clean shape x_0. This expression is defined as:

$$q(x_t|x_0) = \mathcal{N}(x_t; \alpha_t x_0, \sigma_t^2 I) \tag{A.3}$$

where $\alpha_t = \sqrt{\prod_{s=1}^{t}(1-\beta_s)}$ and $\sigma_t = \sqrt{1-\alpha_t^2}$.

The reverse diffusion process is implemented as a second parameterized Marcov process with parameters θ that inverts the forward diffusion, conditioned on the partial shape c:

$$p_\theta(x_{0:T-1}|x_T, c) = \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, c) \tag{A.4}$$

with:

$$p_\theta(x_{t-1}|x_t, c) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, c, t), \rho_t^2 I) \tag{A.5}$$

where the mean $\mu_\theta(x_t, c, t)$ is learned by the network and ρ_t^2 are time step-dependent fixed variances.

To simplify the objective function used to train the model, the mean and variances in eq. A.5 are reformulated as follows:

$$\mu_\theta(x_t, c, t) = \frac{1}{\sqrt{1-\beta_t}}(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t^2}}\epsilon_\theta(x_t, c, t)) \tag{A.6}$$

$$\rho_t^2 = \frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t \tag{A.7}$$

The network in this case takes the noisy input $x_t \sim q(x_t|x_0)$, a diffusion step t and a partial point cloud as a condition c and predicts $\epsilon_\theta(x_t, c, t)$ in A.6. Intuitively, ϵ is the actual noise added to the clean shape to obtain x_t and ϵ_θ is the noise predicted by the model. The simplified loss function is given in eq. A.8.

$$\mathcal{L}(\theta) = \mathbb{E}_{i\sim\mathcal{U}([M]),t\sim\mathcal{U}([T]),\epsilon\sim\mathcal{N}(0,I)} \left\|\epsilon - \epsilon_\theta(\alpha_t x_t^i + \sigma_t\epsilon, t, c^i)\right\|^2 \tag{A.8}$$

where $\mathcal{U}([M])$ represents the uniform distribution over {1, 2,...,M}, $\mathcal{U}([T])$ represents the uniform distribution over {1, 2,...,T}.

At inference, ancestral sampling is performed in an iterative fashion to generate the complete shape. First, x_T is obtained by sampling from $\mathcal{N}(0, I)$ then x_{T-1} through x_1 are iteratively drawn from $p_\theta(x_{t-1}|x_t, c)$ (eq. A.5) until a clean shape x_0 is obtained. Note that we only sample a data pair (x_i, c_i) from the training set, a time step t, and Gaussian noise ϵ at each training step. In other words, the iterative sampling approach is not performed during training.

Appendix B. Additional Qualitative Results

Appendix B.1. Output Diversity for The Same Input

To investigate the stochasticity of the model, we generated different outputs using the same conditioner for a set of cases. Fig. B.15 illustrates example results for two cases.

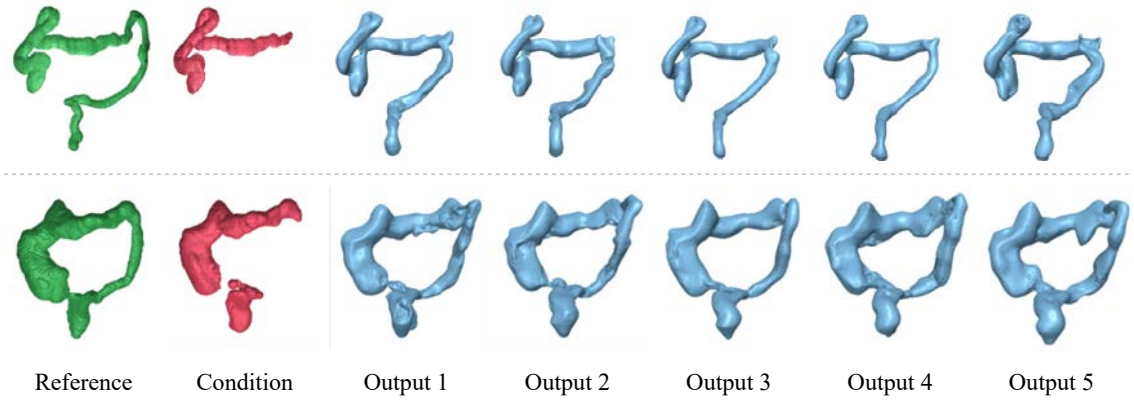


Figure B.15: Examples of different outputs generated using the same condition.

The evaluation revealed that given the same conditioner, the model preserves the global appearance of the outputs and generates shapes that differ in fine details which is consistent with the dynamic filling status of the organ. This observation indicates the fidelity of the model to its input and its stability.

Appendix B.2. Global-Conditioned Generation

To test the nature of the features that the VAE is encoding, we conduct an experiment where we only condition the ResNet-based DDPM on the global latent representations of the erroneous masks whereas we use an unconditioned PVCNN model for the local DDPM. We train the model for 8000 epochs. Examples of the generated shapes can be seen in Fig B.16

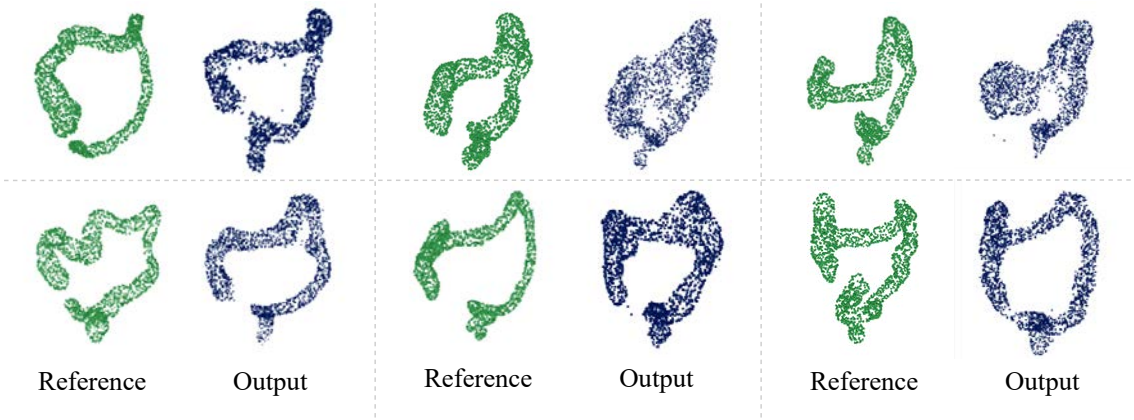


Figure B.16: Example outputs of the model trained using the global conditioning only.

The results show that the generated shapes differ largely in the details such as curvatures, thickness, and point positions compared to the reference but exhibit certain similarities in the overall appearance which confirms that the global latent representations are encoding coarse-level features that express the global appearance of the shape.

Appendix C. TotalSegmentator Refinement

Figure C.17: Examples of TotalSegmentator outputs and their refinement results.

Medical Imaging and Applications

Master Thesis, June 2023

Synthetic Dynamic Contrast Enhanced Breast MRI Generation

Eashrat Jahan Muniya, Dr. Robert Martí

VICOROB research institute, University of Girona, Girona, Spain

Abstract

In recent years, significant advancements have been made in the field of computer vision, particularly in the area of image generation. A notable breakthrough has come in the form of diffusion probabilistic models, which have shown impressive capabilities in generating high-quality images from textual input. These models, such as DALL-E 2, Imagen, and Stable Diffusion, have revolutionized the way we perceive the connection between text and visual data. However, the systematic evaluation of these models in the medical domain, where image data often consists of three-dimensional volumes, remains limited.3D or 2D Synthetic images have significant potential in preserving privacy in artificial intelligence applications and can serve as a means to enhance small datasets. In this study, we explore the utilization of latent diffusion models in synthesizing medical imaging data, specifically focusing on breast imaging in dynamic contrast-enhanced MRIs. Our approach involves generating three-dimensional volumes of pre-contrast MRIs and subsequently synthesizing two-dimensional post-contrast slices. To quantitatively assess the performance of the synthesized images, we incorporate measurements and expert evaluations from medical professionals who rate the quality of the post-contrast synthesized images. Our results demonstrate the effectiveness of diffusion probabilistic models in generating realistic medical imaging data, offering new possibilities for privacy preservation in AI and data augmentation in the medical field.

Keywords:
Breast MRI, Synthetic image generation, Diffusion models, Image to image generation, Breast cancer diagnosis

1. Introduction

Cancer, a pressing and growing global concern, stands as a prominent contributor to morbidity and mortality, currently causing one in six deaths worldwide Ferlay et al. (2018). Within this somber context, breast cancer emerges as a matter of utmost significance. With an annual incidence of over 2 million new cases, breast cancer constituted 11.6% of all cancer cases in 2018, affecting 24.2% of women. This alarming prevalence positions breast cancer as the most frequently diagnosed cancer and the leading cause of mortality (6.6%) among women on a global scale Wild C.P. (2020). Deep learning models have played a crucial role in significant breakthroughs in various domains such as natural language processing and computer vision. These achievements can be attributed to the extensive training of these models with diverse datasets. In addition, the generation of synthetic medical data offers a promising and viable alternative, enabling large-scale research to be conducted effectively while preserving the privacy of the patients Jordon et al. (2022), Wang et al. (2021). This approach allows for meaningful exploration and analysis while mitigating concerns regarding the privacy and security of sensitive data. Some publicly available datasets have grown to contain millions of images and text sentences, contributing to the improved performance of deep neural networks in these domains. Deng et al. (2009). Significant advancements were witnessed in medical image analysis through the utilization of deep neural networks to address various tasks, including segmentation, structure detection, and computer-aided diagnosis. Magnetic Resonance Imaging (MRI) has revolutionized the detection of breast lesions, showcasing remarkable advancements in the field of medical image analysis. These advancements include the development and application of deep neural networks for tasks such as accurate segmentation of breast structures,

precise detection of abnormalities, and computer-aided diagnosis, leading to improved diagnostic accuracy and patient care. Shen et al. (2017). However, one current limitation of medical imaging projects is the lack of availability of large datasets. Additionally, acquiring high-quality breast MRI images poses challenges due to long scan times, motion artifacts, the need for contrast agents, and high acquisition costs. This master's thesis focuses on the development and evaluation of two interconnected methodologies that leverage diffusion models to generate Dynamic Contrast-Enhanced (DCE) breast MRI imaging. DCE breast MRI is a widely used imaging technique that provides valuable information for the diagnosis and characterization of breast lesions. DCE breast MRI offers several advantages over other imaging modalities, such as mammography or ultrasound, as it can detect subtle changes in blood flow and vascular permeability, aiding in the detection and assessment of breast lesions Mann et al. (2019).In DCE breast MRI, the choice of contrast agent and its administration protocol is crucial. Gadolinium-based contrast agents are commonly used, administered intravenously, to capture the enhancement pattern over time. Different imaging sequences, such as T1-weighted or T2-weighted, provide valuable information about lesion morphology and tissue characteristics. Another important fact is the resolution plays a vital role where higher spatial resolution allows the detection of smaller lesions and improves lesion characterization. To illustrate the concepts discussed, in Figure 1 an example of a DCE breast MRI can be visualized. Figure 1 shows a set of images representing the pre-contrast, post-contrast, and contrast uptake curve. Contrast uptake patterns and wash-in/wash-out rates offer valuable insights into the characteristics of a lesion. Kuhl et al. (2005) proposed a widely used classification scheme that categorizes breast masses based on the shape of their contrast-time intensity curves. The classification scheme indicates that breast masses exhibiting a gradual increase, in contrast, uptake (Type I pattern) are more likely to be benign, while masses with rapid contrast uptake and wash-out (Type III) are more likely to be malignant. Lesions displaying a plateau enhancement pattern (Type II) fall in an intermediate category.

The first part of the thesis addresses the generation of high-resolution synthetic pre-contrast 3D breast MRI images using diffusion models. Synthetic pre-contrast 3D breast MRI images are essential in establishing baseline characteristics of breast lesions, providing crucial information for accurate diagnosis and treatment planning. However, acquiring high-quality breast MRI images for every patient is often impractical or resource-intensive. Synthetic imaging techniques offer a promising solution by enabling the generation of realistic and high-fidelity images, bridging the gap between limited imaging resources and the increasing demand for accurate diagnostic tools. Later we focus on addressing

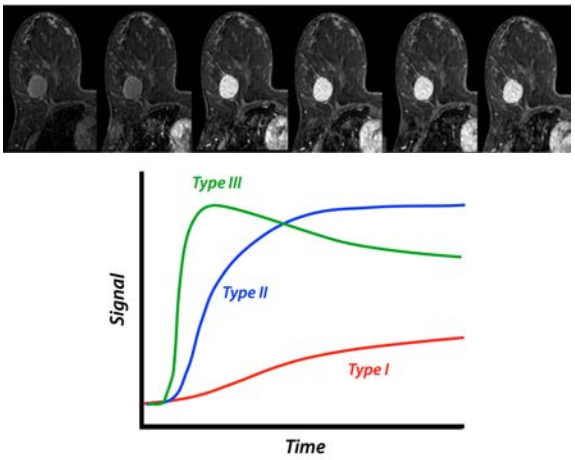


Figure 1: Tumor MR enhancement patterns (Elster).

the need for post-contrast images, which play a crucial role in evaluating the enhancement patterns of breast lesions. Traditional post-contrast imaging requires the administration of contrast agents, which can be invasive and time-consuming. Our objective is to develop a methodology that can generate synthetic post-contrast 2D images from the pre-contrast images, eliminating the need for additional contrast agent administration which is used to enhance the visibility of tumors, lesions, inflammations, and blood vessels. Additionally to have diverse synthetic post-contrast to increase the limited amount of data.

The ultimate goal of this research is to contribute to the field of medical imaging and improve patient care by developing and evaluating the methodologies for generating synthetic pre-contrast 3D breast MRI images and synthetic post-contrast 2D images from the obtained synthetic pre-contrast images. The subsequent chapters of this thesis will provide a detailed explanation of the methodology, experimental setup, results, and discussions of the development and evaluation of the proposed techniques. Additionally, the limitations and future directions of this research will be explored, paving the way for further advancements in breast MRI imaging.

2. Related Work

The success of different types of generative models in various fields has fueled further research and exploration of their potential applications across various creative, entertainment, and data-driven fields. Following this the medical domain has shown persistent interest in harnessing the success of latent diffusion models, recognizing their potential for a diverse range of applications. The scarcity of medical data, particularly in the form of 3D CT and MRI data, has posed a significant limitation in the progress and innovation of medical imaging and analysis techniques. Generative Adversarial Networks (GANs) on the other hand have gained

widespread usage across diverse domains for generating synthetic images both medical and non-medical domains Creswell et al. (2018), Kwon et al. (2019). To mitigate the problem of computational expenses, A hierarchical 3D Generative Adversarial Network (GAN) was proposed by Sun et al. (2022), allowing the generation of realistic 3D thorax CT and brain MRI images with resolutions of up to $256 \times 256 \times 256$ voxels. The approach involved generating a low-resolution version of the image and using it as a reference to generate high-resolution sub-volumes. Still, GANs have the limitation of their instability during training, often failing to converge and struggling to capture the full range of variability in the generated data. This issue, commonly referred to as mode collapse results in generating diverse samples Kodali et al. (2017). Dhariwal and Nichol (2021) showed that diffusion models can achieve image sample quality superior to the current state-of-the-art generative models. But there is a relatively small community of researchers dedicated to exploring diffusion and latent diffusion models for the generation of synthetic 3D medical data. Three-Dimensional Medical Image Synthesis with Denoising Diffusion Probabilistic Models by Dorjsembe et al. (2022) demonstrated how 3D brain images are generated with the help of Diffusion probabilistic models. This work presents the first attempt to investigate the DDPM to enable 3D medical image synthesis. They utilized their approach on 1500 contrast-enhanced anonymized T1 images which is comparatively a small dataset. But their generated samples performed similarly to the GAN-based models with an MS-SSIM Wang et al. (2003) score of 0.8241. There is one publication that has utilized diffusion models in the latent space to generate 3D MRI data using a comprehensive database of brain scansPinaya et al. (2022). They used the UK Biobank datasetSudlow et al. (2015) which has 31740 MRI images. They were able to sample high-quality images with sharp details and realistic textures compared to VAE-GANs, and LSGAN2. In addition to generation diversity, they outperformed the other models with MS-SSIM of 0.65 and 4-G-R-SSIM (Rouse and Hemami (2008),Chen et al. (2006)) of 0.3883. Despite the existence of studies focusing on synthetic brain and other organ generation, the research on synthetic 3D breast MRI generation remains relatively limited. A recent work "Medical Diffusion: Denoising Diffusion Probabilistic Models for 3D Medical Image Generation" by Khader et al. (2022) was the only available work found for 3D breast MRI synthesis. To encode images into a meaningful latent representation, they have adapted VQ-GAN's Esser et al. (2021) latent representation. Along with other medical data, The authors of the study evaluated their model by utilizing a breast cancer MRI dataset Saha et al. (2018) as the same dataset used in our project which aligns with the dataset employed in our project. Unlike synthetic medical data generation, there are several works out there that demonstrate and facilitate the need for contrast synthesis. Grant-Jacob et al. (2021) explored the transformation in magnetic resonance imaging via Deep learning using data from a single asymptomatic patient where out of various image-to-image translation models pix2pix Isola et al. (2017) yielded better results. Another work on the synthesis of post-contrast T1-weighted MRI for tumor response assessment in neuro-oncology Preetha et al. (2021) reported that cGAN based on pix2pix achieved SSIM score of 0.818 which is higher than 3D CNN based UNet. Costa et al. (2017) proposed an approach that learns to generate eye fundus images through the data. The authors matched real eye fundus images with their proper vessel tree and trained them to train retinal vessel segmentation images. Then, the model was trained the interpretation from the vessel tree to a generated retinal image.

Overall, In the context of breast imaging, there is not much available prior work for synthesis. To address this gap and achieve our objective, we propose a synthetic breast MRI image generation process utilizing established latent diffusion models. Our study focuses on investigating the influence of various parameters and settings on the quality and performance of synthesized images. We particularly emphasize the application of latent diffusion models, including autoencoders with VQ-VAE/KL-regularization. Additionally, we adopt the widely acknowledged pix2pix architecture for the post-contrast generation of 2D breast MRI slices.

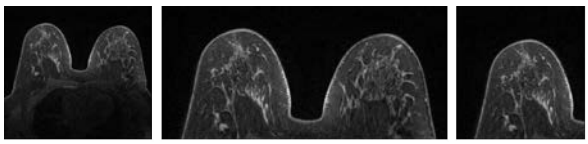
3. Material and methods

3.1. Data Acquisition

The breast cancer MRI dataset utilized in this study is sourced from the publicly available dataset Saha et al. (2018). The dataset is a retrospective collection from a single institution and comprises 922 patients diagnosed with biopsy-confirmed invasive breast cancer. The data covers a span of over a decade, providing a comprehensive representation of cases during that period. The dataset was curated to support research in the field of breast cancer diagnosis and characterization using MRI imaging. Pre-operative dynamic contrast-enhanced (DCE)-MRI data were obtained from Picture Archiving and Communication Systems (PACS) and de-identified for release in The Cancer Imaging Archive (TCIA). The dataset consists of axial breast MRI images acquired using 1.5T or 3T scanners with patients in the prone position. The DICOM format contains several MRI sequences, including a non-fat saturated T1-weighted sequence, a fat-saturated gradient echo T1-weighted pre-contrast sequence, and typically three to four post-contrast sequences. The DCE-MRI images were annotated by radiologists, indicating the locations of lesions within the breast.

3.2. Data Pre-processing

The breast cancer dataset was first preprocessed by stacking all the DICOM sequences of each patient into a single nifti volume. This merging was performed while preserving the original spacing and orientation of the sequences. Afterwards, all images were resampled to a voxel spacing according to the first volume (0.803 mm, 0.803 mm, 1 mm) and then using the corresponding annotated ROI that outlined the breast to crop out a region of interest of only the breast out of the chest. Different orientation conventions commonly used in the context of three-dimensional (3D) MRI images RAI (Right-Anterior-Inferior), LPI (Left-Posterior-Inferior) were found. These views represent different coordinate systems that define the spatial orientation of the image slices within the 3D volume. As using a consistent orientation across all images ensures that the information is aligned consistently for the diffusion model, all the images were reoriented to the RAI orientation. The preprocessing steps described in this work were derived from the methods proposed by Khaled et al. (2021) in their study, where they utilized the same dataset. The images were then split into two halves, such that the left and the right breast were on separate images. By splitting the volumes the dataset was a total of 1844 separate volumes in the end. In Figure 2, the breast cropping process for patient 1 is illustrated.



(a) Whole Breast (b) Cropped Breast (c) Splitted Breast

Figure 2: Breast ROI cropping and splitting visualization for Patient 1 on the Middle Slice along Axial Plane.

Finally, the images were resized to a uniform shape of 64x64x32, 96x96x64, and 128x128x96 voxels based on the different experiments. All the images were min-max normalized to the range between -1 and 1. The preprocessing steps were done for the pre to post-contrast synthesis involved extracting the middle slice consisting of the lesions along a specified z-axis(axial) figure 3 from each sequence of both pre and post-contrast images. The middle slice in MRI is often considered important in medical imaging, particularly in the context of lesion analysis. This is because the middle slice is typically chosen to represent the central portion of the anatomy being imaged, providing a representative view of the structures and abnormalities within the region of interest. Subsequently, the extracted slice was cropped to isolate the breast region, if present, thereby focusing on that specific area of interest. The normalization process was followed by global normalization.

Global Normalization: In global normalization, pre-contrast sequences are normalized together, and the intensity values of the voxels from the post-contrast sequences are normalized together. The intensity values of the voxels across the pre-contrast and post-contrast sequences are adjusted to their respective common scale, Finally, the resulting cropped slices were saved as PNG files. In the context of using the pix2pix

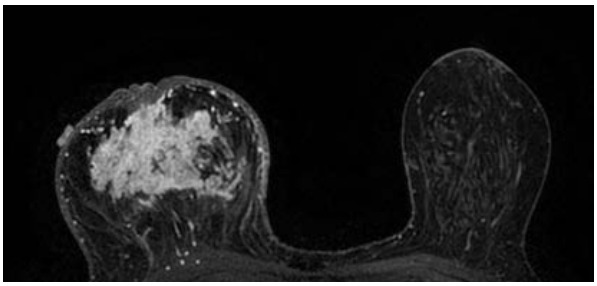


Figure 3: Sample of an Extracted middle slice of post-contrast with visibly enhanced regions.

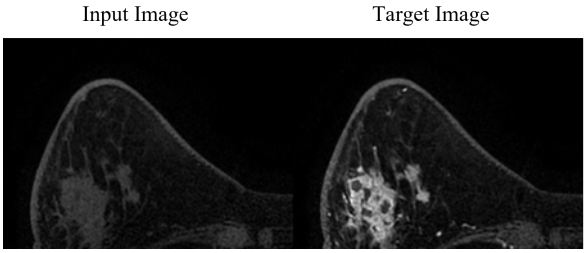


Figure 4: : Sample training Pair of Pre and Post-contrast(Global Normalization)

architecture Isola et al. (2017) for contrast synthesis, the input images consist of a set of pre-contrast images and their corresponding post-contrast images. To facilitate the training and translation process, these paired images are arranged in a specific format where the pre-contrast image is stacked adjacent to the corresponding post-contrast image. The resulting combined image exhibits a side-by-side configuration, where the left part of the combined image represents the pre-contrast input image, and the right part represents the corresponding post-contrast target image 4. As a result, the model receives a pair of real input and real target aiming to generate a fake target. During training the data was augmented by applying the ColorJitter transformation, with a specific emphasis on contrast. Contrast - refers to the difference in brightness between the image's lightest and darkest areas. This data augmentation technique introduced random color variations to the images, enhancing the contrast component. During training, all the images were resized to the size of 256x256 pixels.

3.3. Architechture

DCE MRI synthetic Generation utilizes the Latent Diffusion Model (LDM) Rombach et al. (2022) for generating synthetic pre-contrast 3D MRI images. The focus is on generating both pre and post-contrast images.

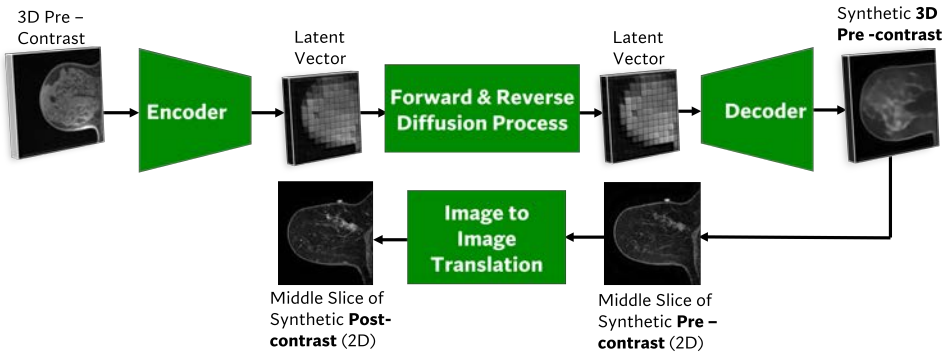


Figure 5: Schematic concept of the synthetic Dynamic Contrast breast MRI Generation.

The middle slices of MRI images are considered crucial for capturing important details about the lesion. Extracted pre and post-contrast 2D middle slices are paired for image-to-image translation. This approach allows the generation of synthetic post-contrast images. Overall, the proposed architecture combines the Latent Diffusion Model, and image-to-image translation to generate high-quality synthetic 3D pre-contrast and post-contrast 2D MRI images, effectively capturing the desired information and enhancing the representation of contrast-enhanced regions. Figure 5 shows the architecture of the proposed method. data.

3.3.1. Latent Diffusion Model

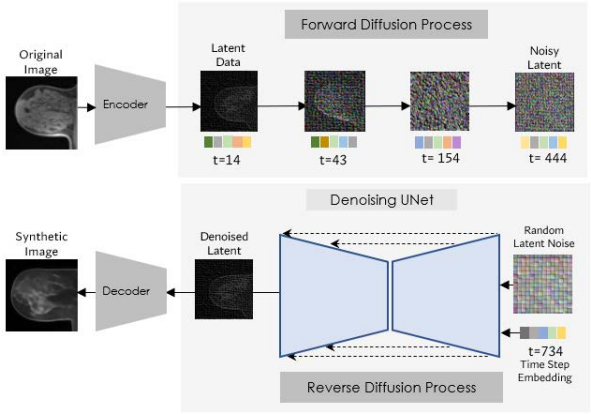


Figure 6: Latent Diffusion model Architecture

In order to generate 3D high-resolution images LDM was used, which is a two-step approach. Firstly, an encoder model is trained to encode the input images, producing a latent representation. Then, the diffusion probabilistic model is trained using this latent representation to generate synthetic images with its generative properties. This process involves training an encoder to extract meaningful information from the input data, which is then utilized by the diffusion model for image synthesis. During the encoding of the image, it adds additional channel numbers that represent the number of channels or feature maps present in the input data. It represents the dimensionality of the input data in the channel axis. MONAI (Medical Open Network for AI) framework developed by the MONAI community Consortium (2022) was used for implementing the diffusion and Latent diffusion model in our project. The encoder and diffusion components from MONAI provided a robust foundation for our research, enabling us to effectively model and generate synthetic medical images. To obtain the latent representation 2 encoder architectures were explored. Figure 6 shows a generic architecture of LDM used for this project. In the following, background information on the Vector Quantised-Variational Autoencoder Van Den Oord et al. (2017), KL encoder Pinaya et al. (2022), and Denoising diffusion probabilistic model Sohl-Dickstein et al. (2015) are given.

VQ-VAE: The VQVAE (Vector Quantised-Variational Autoencoder) is a generative model that combines concepts from variational autoencoders (VAEs) and vector quantization. It consists of an encoder network, a decoder network, and a codebook of discrete latent vectors(vector quantizer). The goal of the VQVAE is to learn a compressed representation of input data that captures the salient features. The VQVAE is trained using a combination of reconstruction loss and quantization loss. The implementation of the VQVAE was done in MONAI framework based on Van Den Oord et al. (2017). Different numbers of embedding dimensions were explored in order to get the proper reconstruction of the images including 3, 8, 16, and 32. Figure 7 shows the number of channels and the containing morphological attributes of an input image after encoding it with an embedding dim of 16 along each channel. During training the VQ-VAE model for reconstruction showed great performance with embedding dim greater than 16. While integrating with the diffusion model, it collapsed and gave nan values for the prediction. The prediction showed improvement with an embedding dimension of 3; however, the reconstruction was poor, and the prediction lacked coherence or logical consistency with the diffusion model. Thus VQ-VAE was not part of the final pipeline

and its integration was left for future work.

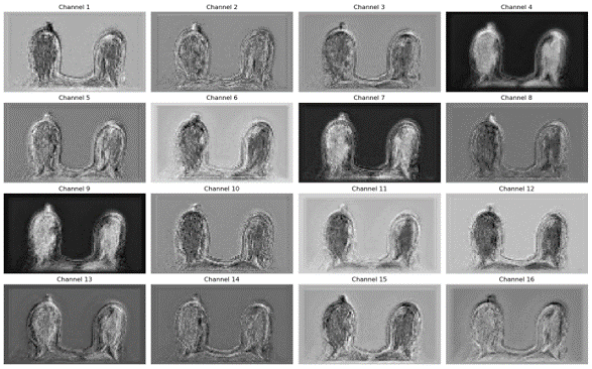


Figure 7: Sample encoded image with embedding dimension of 16. (i.e if the image is 128x128x96,the embedded image will be of 16x32x32x24)

KL Encoder: The autoencoder model is trained in an adversarial manner following Esser et al. (2021). The implementation of the KL encoder is an adaptation of Pinaya et al. (2022), where they use LDM to generate synthetic Brain MRI. The AutoencoderKL model consists of an encoder and a decoder. The encoder takes input data and maps it to a lower-dimensional latent space representation, while the decoder reconstructs the original input data from the latent space representation. The compression model was an essential step to allow us to scale to high-resolution medical images. The AutoencoderKL model uses multiple loss functions for training with a combination of L1 loss, perceptual loss Zhang et al. (2018), a patch-based adversarial objective Wang et al. (2018), and a KL regularization of the latent space. The encoder maps the breast image to a latent representation with a size of $24 \times 24 \times 16$, $32 \times 32 \times 24$ depending on the different-sized experiment. The channel dimension of the latent embedding is set to 3, indicating that each atomic element in the latent space has a vector of three values. In a study conducted by Khader et al. (2022). on medical diffusion, it was observed that using a smaller compression factor of 4 resulted in a more precise reconstruction of anatomical features. Thus, in our experiments, we kept the compression factor of 4 (i.e., images of size 1287x128x96 have a latent dimension of 32x32x24)

Diffusion model: Diffusion models, as proposed by Sohl-Dickstein et al. (2015), are probabilistic models that refer to a parameterized Markov chain specifically designed to learn a data distribution, denoted as $p(x)$. Its objective is to generate samples that closely match the given data distribution $p(x)$ within a finite time frame. These models achieve this by iteratively removing noise from a normally distributed variable. In essence, the learning process involves reversing a fixed Markov Chain of length T (Pinaya et al., 2022).

- Forward Diffusion Process: In the forward diffusion process we take an MRI image x_0 and continuously destroy the structural integrity by adding Gaussian noise for increasing timesteps T such that they move out or move away from their existing subspace which is a normal/gaussian distribution as shown in Figure 8. Typically done using a fixed linear variance scheduler. So, given the data at time step $t-1$, $q(x_t|x_t-1)$ known as forward diffusion kernel, represents the normal distribution of the data at timestep t. The distribution q in the forward diffusion process is defined as Markov Chain given by equation 1.

$$q(x_t|x_{t-1}, x_0) = \mathcal{N}(x_t;\ \sqrt{\mathbf{1-\beta_t}}\, \boldsymbol{x_{t-1}}, \beta_t I) \quad (1)$$

- Reverse Diffusion process: The idea is to reverse the forward diffusion process can be visualized in Figure 8. The reverse process is modeled as another Markov chain, aiming to reconstruct the original input x_0 from the noisy version. Through iterative learning, this reverse chain learns to undo the effects of the noise and restore the MRI image to its original form. So, given the data at time step t, it models the probability of the previous data point $x_t - 1$ so that the data distribution $p(x_t - 1|x_t)$, known as reverse diffusion kernel can be inferred to recover the original input from the noisy data.

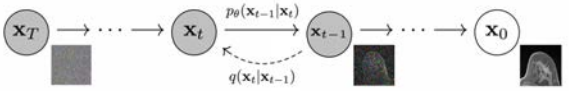


Figure 8: The directed Diffusion graphical model from Ho et al. (2020).

- Adding Noise: The DDPM paper describes a corruption process that adds a small amount of noise for every 'timestep'. Given x_{t-1} for some timestep, we can get the next (slightly more noisy) version x_t with equation 1. Where, we take x_{t-1}, scale it by $\sqrt{1-\beta_t}$ and add noise scaled by β_t. This β is defined for every t according to some schedule and determines how much noise is added per timestep. Now, we don't necessarily want to do this operation 500 times to get x_{500} so we have another formula to get x_t for any t given x_0 (Hug). Initially, the noisy x is mostly x ($\sqrt{\bar{\alpha}_t} = 1$) but over time the contribution of x drops, and the noise component can be visualized in Figure 9.

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t;\ \sqrt{\bar{\alpha}_t}\mathbf{x}_0,\ \sqrt{(1-\bar{\alpha}_t)}\mathbf{I}) \quad (2)$$

The neural network used to model the noise is typically chosen to be a U-NetRonneberger et al. (2015), Khader et al. (2022). In order to support 3D data, 2D convolutions were replaced by 3D convolutions. The

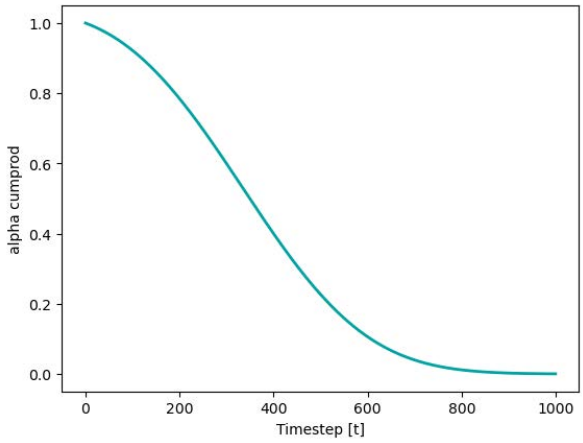


Figure 9: Noising across different timesteps.

implementation of the U-net model was accomplished using MONAI. Where the Noise Scheduler component was incorporated into the diffusion model. We utilized a DDPMScheduler, which allows us to configure the diffusion process with 1000 timesteps and a `scaled_linear` profile for the beta values. This particular profile, proposed in Rombach et al. (2022) work on "High-Resolution Image Synthesis with Latent Diffusion Models," yielded superior outcomes compared to the linear profile originally introduced in the DDPM's paper. We set the parameters `beta_start` and `beta_end` to define the range of beta values, which play a pivotal role in determining the intensity of noise added to the images.

To generate synthetic images, first, the autoencoder model was trained on the whole dataset with the desired size in our case the 2 different input sizes 128x128x96, and 96x96x64 voxels. The diffusion model expects the input to be normalized in the range of -1, 1 (Ho et al., 2020) so, the latent representation was also closer to this range. Once the encoder model is trained and knows how to encode and decode, this was used to train the diffusion model to generate the 3D synthetic volumes.

3.3.2. Pre to Post Contrast Synthesis

To generate fake post-contrast breast MRI images the architecture of Isola et al. (2017), "Image-to-Image Translation with Conditional Adversarial Networks" was followed. The pix2pix model is mainly based on the concept of cGAN (conditional generative adversarial network), which combines a generator network and a discriminator network. The objective of the pix2pix model is to learn a mapping between a source image x and a random noisy image z to generate the corresponding target image y, denoted as $x, z \rightarrow y$. The discriminator network is trained to distinguish between real target images y given the source image x and fake target images generated by the generator. The objective function of the pix2pix model represents the loss or error that measures the difference between the generated images and the real target images, guiding the training process to improve the quality and realism of the generated outputs as follows.

$$L_{Pix2Pix}(G, D) = E_{x,y}[\log D(x, y)] + E_{x,z}[\log(1 - D(x, G(x, z)))] \quad (3)$$

In the following, background information on the Generator and the Discriminator of pix2pix is given.

- U-net Generator : The generator of the pix2pix cGAN is a modified U-Net Ronneberger et al. (2015). A U-Net consists of an encoder (downsampler) and a decoder (upsampler). During the downsampling phase, spatial information is progressively extracted and passed from one convolutional block to the next, culminating in the bottleneck region. Upsampling, on the other hand, begins from the bottleneck and involves transpose convolutional blocks that expand the information while incorporating details from the corresponding downsampling blocks. This concatenation of information enables the network to learn and generate a more accurate output by leveraging the combined knowledge from different scales of the input data. The U-Net architecture allows for capturing both low-level and high-level features in the image translation process. It helps to preserve the spatial information and improve the overall performance.
- Markovian Discriminator Architecture: The discriminator network in pix2pix employs a PatchGAN architecture, which performs image classification at the patch level. Instead of classifying the entire image as real or fake, it assesses the authenticity of the $N \times N$ image patch. The discriminator is run convolutionally across the image with 70×70 patches, averaging all responses to provide the final output. The discriminator has fewer parameters compared to the generator, it is effectively faster.

The generator and discriminator are trained in an adversarial manner. The generator aims to minimize the adversarial loss, which encourages the generated images to be classified as real by the discriminator as illustrated in Figure 10. Additionally, a pixel-wise loss L1 is used to enforce pixel-level similarity between the generated and real images. The L1 loss with a λ factor of 100 is added to the cross entropy loss. According to Isola et al. (2017) this yields better results. The combined loss guides the learning process and facilitates the generation of realistic synthetic post-contrast MRI images.

3.4. Experiments

To use the latent representation of VQ-VAE different numbers of latent channels were used for the reconstruction and the in this case the larger number the better the

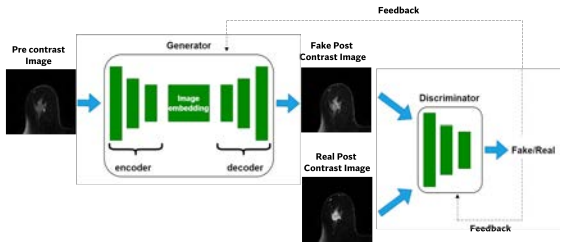


Figure 10: Diagram illustrating the process for training the Pix2Pix network

reconstruction was. A latent channel of 16 yielded better results than the one with 8 or 3. For LDM both encoder and diffusion models are trained on NVIDIA A40 with 46 GB graphics processing unit (GPU) RAM card with a batch size of 4 for 150K iterations. To address memory limitations and enhance the training process in the diffusion model, we employed gradient accumulation techniques by accumulating gradients over multiple smaller batches. Both models took approximately 4 days to finish training. The KL encoder was trained in 2 different data sizes 96x96x64, and 128x128x96 voxels. It is important to emphasize that the autoencoder must undergo training using data that matches the same size as the data it will be used within the diffusion model. Given that contrast synthesis operates in a two-dimensional context, the approach of employing available pre-trained weights from the pix2pix model was conceived. But as those models are trained for specific pair of natural images it wasn't feasible to use them for medical imaging. The Pix2Pix was trained using an NVIDIA A30 with a 24 GB graphics processing unit (GPU) RAM card for approximately 2 hours with a batch size of 16,8 and 4 for 500 epochs.

3.5. Experimental Setup

The software implementation for this project utilized various tools and technologies. The core framework employed was MONAI (Medical Open Network for AI) version 1.1.0, which is a Python-based open-source framework specifically designed for medical imaging tasks. The implementation was done in Python version 3.8.16. Visual Studio Code (VSCode) was used as the IDE for development. The implementation was carried out on a Linux operating system. The NVIDIA A40 46GB GPU and NVIDIA A30 24GB GPU were used for accelerated computations during training and inference.

4. Results

Even though we do not have any prior work to compare any metrics to evaluate the synthetically generated 3D volumes we have used multi-scale structural similarity metric (MS-SSIM) wang2003multiscale to evaluate the diversity of our LDM model. MS-SSIM is a more advanced form of SSIM, performed at multiple scales through a multi-step downsampling process. We averaged 50 synthetic sample pairs of the same dataset. Higher MS-SSIM scores indicate the similarity between the generated synthetic images is similar. Conversely, lower MS-SSIM scores suggest a lower resemblance between the synthetic images. Thus a lower MS-SSIM demonstrates the model is capable of generating more diverse images. In Mao et al. (2017) the GAN model, with its high MS-SSIM score of 0.999, lacks the capability to generate diverse images. Consequently, the synthetic images produced by the GAN model often appear identical. Table 1 shows the MS-SSIM scores of LDM for experimented sizes where we have lower MS-SSIM scores similar to the original images.

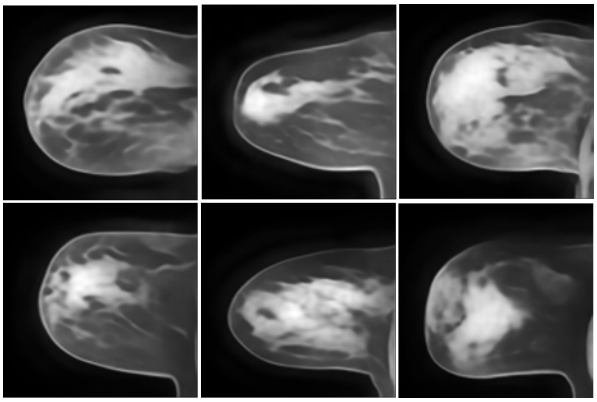


Figure 11: Visualization of Middle Axial slice of Generated 128x128x96 (height, width, depth) synthetic MRI volumes with LDM.

Table 1: MS-SSIM scores for synthetic images with LDM and real images.

Modality (Height, width, depth)	**Image Size**	**MS-SSIM**
LDM	96x96x64	0.54 ± 0.08
LDM	128x128x96	0.57 ± 0.15
Real Images	128x128x96	0.49 ± 0.10

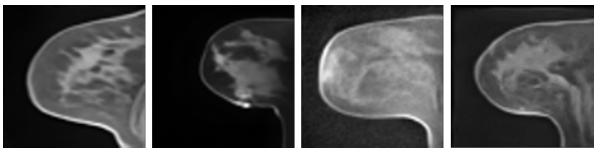


Figure 12: Visualization of Middle Axial slice of Generated 64x64x32 (height, width, depth) synthetic MRI volumes with DDPM.

For the contrast synthesis, we calculated structural similarity (SSIM) (Yao et al., 2021) between the synthetic post-contrast and its ground truth expressed in equation 4.

$$\mathrm{SSIM}(I,G) = \frac{(2\mu_I\mu_G + C_1)(2\sigma_{IG} + C_2)}{(\mu_I^2 + \mu_G^2 + C_1)(\sigma_I^2 + \sigma_G^2 + C_2)} \quad (4)$$

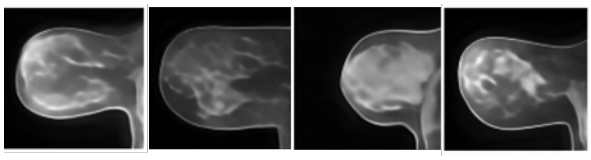


Figure 13: Visualization of Middle Axial slice of Generated 96x96x64 (height, width, depth) synthetic MRI volumes with LDM.

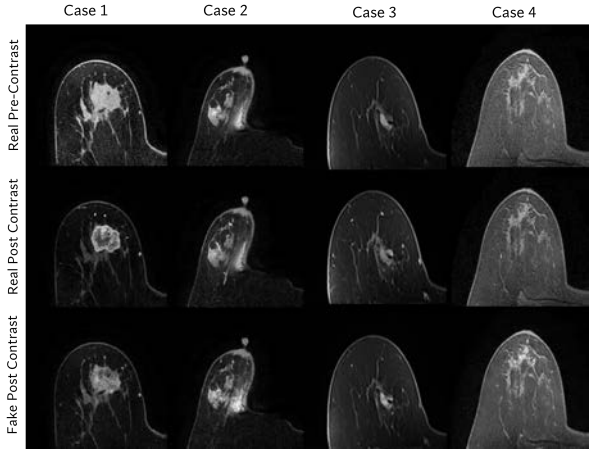


Figure 14: Real and generated post contrast with their corresponding pre Contrast.

In addition, to quantify the performance of the generated images we calculate the Peak Signal-to-Noise Ratio (PSNR) for all the generated images. The PSNR is a metric that quantifies the quality of an image by measuring the ratio of the maximum possible signal power to the power of the noise in the image, defined as

$$PSNR = 10log_{10}\left(\frac{max^2(I,G)}{\frac{1}{N\times M}\Sigma_{M,N}(I(m,n)-G(m,n))^2}\right) \quad (5)$$

where N and M represent the total number of rows and columns of pixels, and m and n represent the pixels in each row and column respectively, the expression 'max(I, G)' refers to the maximum intensity value between the actual ground-truth image I and the generated image G. A greater number represents higher accuracy of the generated image. In Table **??** the PSNR and SSIM scores can be seen.

Table 2: Comparison of SSIM and PSNR values for different modalities

Modality	SSIM	PSNR (dB)
Pix2pix	0.76	26.49
Pix2pix + Data augmentation	0.80	29.53

To further quantify the results in terms of diversity and if the generated post-contrast images can take the spot of the real post-contrast we evaluated the images by a radiologist with 30 years of experience. The expert was shown 30 pairs of Real and fake post-contrast images from the test set and was told to identify the real post-contrast from them. The evaluation consisted of 3 quotas as part of the evaluation: 1)Realistic only minor unrealistic errors, 2) Can't differentiate if fake or real, and 3)Not realistic. Out of the 30 image pairs the radiologist considered 9 fake images as real and for 10 pairs it was not decidable for either fake or real Figure 15 visualizes assessment. Generated synthetic post-contrast images from the test set are illustrated in Figure 14.

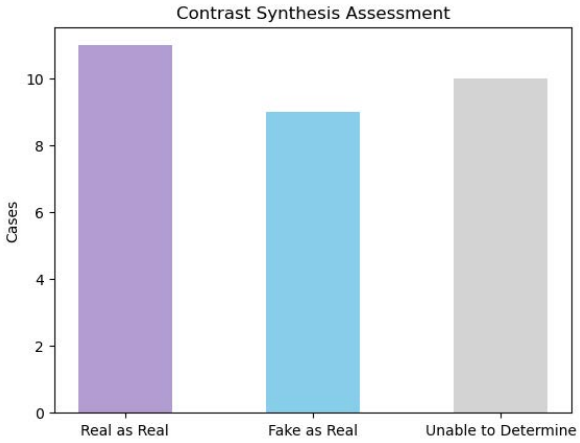


Figure 15: Quantitative evaluation of the image synthesis for 30 test cases.

5. Discussion

In our project, we encountered several challenges that affected the performance of the LDM. These challenges were primarily attributed to limitations in data availability and resizing techniques employed during preprocessing. Training the model with a very small dataset was a major challenge that hindered the encoder models' ability to effectively understand the patterns in the data for encoding. The encoded representations lacked accuracy and couldn't capture the complex details found in the original images16.

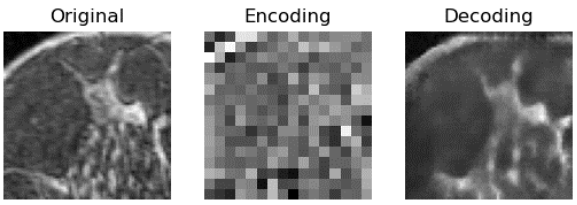


Figure 16: Recontruction of the axial plane with KL encoder

To facilitate training with the available computational resources the images were resized to a smaller volume from $64 \times 64 \times 32$ upto $128 \times 128 \times 96$ voxels. However, this downsizing process introduced its own set of issues. Firstly, downsampling the images caused a loss of essential finer details, leading to a noticeable degradation in the quality of the generated images. The re-

sizing method, as opposed to cropping, further exacerbated the anatomical loss of the images, resulting in a significant impact on the overall image fidelity. To address these limitations, it would be beneficial to consider cropping the images to their desired size instead of resizing them. This approach would preserve more information and resolution, enabling both the encoder and the diffusion model to better learn the underlying data distribution. By mitigating the loss of essential details caused by resizing, the resulting images are expected to exhibit improved reconstruction quality. In (Pinaya et al., 2022) this autoencoder worked very well in terms of reconstruction having the benefit of a large dataset and optimal size while preserving the original resolution. The VQ-VAE reconstruction exhibited realistic reconstruction compared to the Kl encoder 17. However, it was observed that the latent distribution of the VQ-VAE did not align with the expected distribution of the diffusion model. As a result, the VQ-VAE failed to effectively learn during training and provided unrealistic predictions.

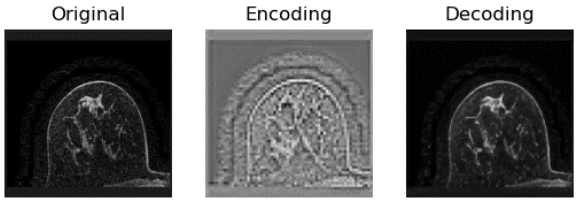


Figure 17: Recontruction of axial plane with VQ-VAE

The images generated with DDPM showed finer and more realistic details than LDM in this case, still, it was not possible to generate the images on a bigger scale as DDPM is highly computationally expensive. One recent work Khader et al. (2022) that takes the latent representation of the VQ GAN for breast with the same dataset provided better reconstruction and due to computational complexity and time limitation. One of the major difficulties encountered in the post-contrast synthesis task was the extensive similarity observed between the pre-contrast and post-contrast images. The goal is to ensure that the similarity between the generated (fake) post-contrast images and the real post-contrast images is higher compared to the similarity between the original pre-contrast and post-contrast images. In Figure 18 we can see that the peak signal-to-noise ratio between pre-contrast and post-contrast and fake post-contrast and real post-contrast is quite close but the fake post vs real post is higher. In addition to, access to a small set of data for training a GAN, this model was performed to capture the similar presence of contrastive areas. The performance improved with data augmentation. Another thing to mention in some cases the fake image exhibits contrast in different areas of the image which clearly didn't show similarity with the ground truth. But during the visual assessment with the radiologist, some cases which were different from the actual post-contrast considered as the real ones. This again can be considered as a benefit of this synthetic generation where it had the ability to generate diverse images while still being in the context.

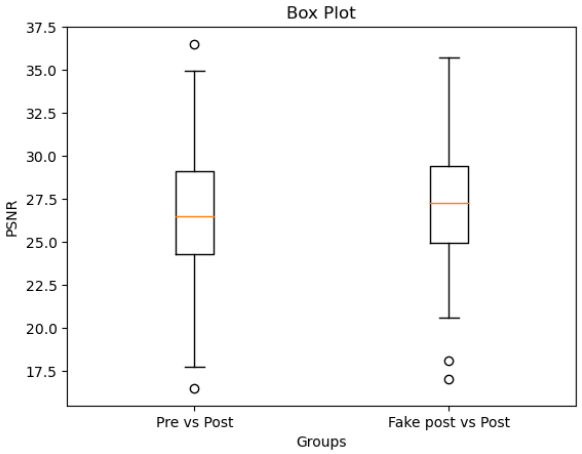


Figure 18: Box plot for PSNR score between pre and real post-contrast and fake and real post-contrast.

The development of synthetic DCE breast MRI generation involves combining two methods, with the intention of utilizing synthetic pre-contrast images to generate synthetic post-contrast images. However, this integration proved challenging due to the smaller size and lack of anatomical detail in the synthetically generated 3D images and left for future work.

5.1. Future Work

Due to limited time and computational resources, encoders with larger-sized images were not possible. To address this issue, we plan to employ mixed precision techniques and gradient checkpointing to optimize the training process and enable the use of larger images. Additionally, we aim to enhance the realism of the generated MRI volumes by experimenting with different parameter settings. Furthermore, we intend to extend our methodology to fully incorporate 3D image-to-image translation techniques, allowing us to work with 3D volumes for more comprehensive transformations. Moreover, we will explore the implementation of the latent representation from the VQ GAN to improve reconstruction quality and make a comparative analysis with existing approaches.

6. Conclusions

In this paper, we explored the implementation of the synthetic breast MRI image generation process using existing diffusion models with an emphasis on latent diffusion models(encoders in particular VQ-VAE/ KL-Encoder) with the extension of contrast synthesis where

we investigated the impact of different parameters and settings on the synthesized images' quality and performance. Despite the limitations posed by the existing literature, dataset size, and computational resources, our method was able to yield fruitful results and successfully generate synthetic 3D and 2D MRI images. The generated images showcased promising potential in capturing relevant anatomical structures and contextual information. While further improvements are necessary to address the challenges we encountered, our approach demonstrates the feasibility and value of synthetic image generation in the field of medical imaging. This opens up possibilities for future research and development in improving the quality and applicability of synthetic MRI images for various clinical applications.

Acknowledgments

First of all, I would like to thank my supervisor Professor Dr. Robert Martí for giving me the opportunity to conduct my thesis on this amazing and interesting topic. Without his constant guidance and support, it would not have been possible. I would also like to thank my MAIA master colleague Ricardo Montoya del for his assistance in order to get started. Thank you MAIA(Medical Imaging and Application) for all these amazing courses. I would also like to thank all my professors for giving me this wonderful educational experience. Thanks to all my classmates for being there and making it a lot easier and memorable. Lastly, I would like to express my heartiest gratitude to the almighty and my family for letting me make it happen.

References

, . Diffusion Model Classhugging face diffusion models course. https://github.com/huggingface/diffusion-models-class. Accessed: 2023-03-4.

Chen, G.H., Yang, C.L., Xie, S.L., 2006. Gradient-based structural similarity for image quality assessment, in: 2006 international conference on image processing, IEEE. pp. 2929–2932.

Consortium, M., 2022. Monai: Medical open network for ai. URL: https://doi.org/10.5281/zenodo.7459814, doi:10.5281/zenodo.7459814. If you use this software, please cite it using these metadata.

Costa, P., Galdran, A., Meyer, M.I., Abramoff, M.D., Niemeijer, M., Mendonça, A.M., Campilho, A., 2017. Towards adversarial retinal image synthesis. arXiv preprint arXiv:1701.08974 .

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A., 2018. Generative adversarial networks: An overview. IEEE signal processing magazine 35, 53–65.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database , 248–255doi:10.1109/CVPR.2009.5206848.

Dhariwal, P., Nichol, A., 2021. Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems 34, 8780–8794.

Dorjsembe, Z., Odonchimed, S., Xiao, F., 2022. Three-dimensional medical image synthesis with denoising diffusion probabilistic models, in: Medical Imaging with Deep Learning.

Elster, A.D., . Questions and answers in mri. URL: https://mriquestions.com/breast-dce.html.

Esser, P., Rombach, R., Ommer, B., 2021. Taming transformers for high-resolution image synthesis, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12873–12883.

Ferlay, J., Ervik, M., Lam, F., Colombet, M., Mery, L., Piñeros, M., Znaor, A., Soerjomataram, I., Bray, F., 2018. Global cancer observatory: cancer today. Lyon, France: international agency for research on cancer 3, 2019.

Grant-Jacob, J.A., Everitt, C., Eason, R.W., King, L.J., Mills, B., 2021. Exploring sequence transformation in magnetic resonance imaging via deep learning using data from a single asymptomatic patient. Journal of Physics Communications 5, 095015.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems 33, 6840–6851.

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks , 1125–1134.

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S.N., Weller, A., 2022. Synthetic data–what, why and how? arXiv preprint arXiv:2205.03257 .

Khader, F., Mueller-Franzes, G., Arasteh, S.T., Han, T., Haarburger, C., Schulze-Hagen, M., Schad, P., Engelhardt, S., Baessler, B., Foersch, S., et al., 2022. Medical diffusion–denoising diffusion probabilistic models for 3d medical image generation. arXiv preprint arXiv:2211.03364 .

Khaled, R., Vidal, J., Martí, R., 2021. Deep learning based segmentation of breast lesions in dce-mri, in: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I, Springer. pp. 417–430.

Kodali, N., Abernethy, J., Hays, J., Kira, Z., 2017. On convergence and stability of gans. arXiv preprint arXiv:1705.07215 .

Kuhl, C.K., Schild, H.H., Morakkabati, N., 2005. Dynamic bilateral contrast-enhanced mr imaging of the breast: trade-off between spatial and temporal resolution. Radiology 236, 789–800.

Kwon, G., Han, C., Kim, D.s., 2019. Generation of 3d brain mri using auto-encoding generative adversarial networks, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22, Springer. pp. 118–126.

Mann, R.M., Cho, N., Moy, L., 2019. Breast mri: state of the art. Radiology 292, 520–536.

Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S., 2017. Least squares generative adversarial networks, in: Proceedings of the IEEE international conference on computer vision, pp. 2794–2802.

Pinaya, W.H., Tudosiu, P.D., Dafflon, J., Da Costa, P.F., Fernandez, V., Nachev, P., Ourselin, S., Cardoso, M.J., 2022. Brain imaging generation with latent diffusion models, in: Deep Generative Models: Second MICCAI Workshop, DGM4MICCAI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings, Springer. pp. 117–126.

Preetha, C.J., Meredig, H., Brugnara, G., Mahmutoglu, M.A., Foltyn, M., Isensee, F., Kessler, T., Pflüger, I., Schell, M., Neuberger, U., et al., 2021. Deep-learning-based synthesis of post-contrast t1-weighted mri for tumour response assessment in neuro-oncology: a multicentre, retrospective cohort study. The Lancet Digital Health 3, e784–e794.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer. pp. 234–241.

Rouse, D.M., Hemami, S.S., 2008. Analyzing the role of visual structure in the recognition of natural image content with multi-scale ssim, in: Human vision and electronic imaging XIII, SPIE. pp. 410–423.

Saha, A., Harowicz, M.R., Grimm, L.J., Kim, C.E., Ghate, S.V.,

Walsh, R., Mazurowski, M.A., 2018. A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 dce-mri features. British Journal of Cancer 119, 508–516. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6134102/, doi:10.1038/s41416-018-0201-5.

Shen, D., Wu, G., Suk, H.I., 2017. Deep learning in medical image analysis. Annual review of biomedical engineering 19, 221–248.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S., 2015. Deep unsupervised learning using nonequilibrium thermodynamics, in: International Conference on Machine Learning, PMLR. pp. 2256–2265.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al., 2015. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine 12, e1001779.

Sun, L., Chen, J., Xu, Y., Gong, M., Yu, K., Batmanghelich, K., 2022. Hierarchical amortized gan for 3d high resolution medical image synthesis. IEEE journal of biomedical and health informatics 26, 3966–3975.

Van Den Oord, A., Vinyals, O., et al., 2017. Neural discrete representation learning. Advances in neural information processing systems 30.

Wang, T., Lei, Y., Fu, Y., Wynne, J.F., Curran, W.J., Liu, T., Yang, X., 2021. A review on medical imaging synthesis using deep learning and its clinical applications. Journal of applied clinical medical physics 22, 11–36.

Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B., 2018. High-resolution image synthesis and semantic manipulation with conditional gans, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8798–8807.

Wang, Z., Simoncelli, E.P., Bovik, A.C., 2003. Multiscale structural similarity for image quality assessment, in: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, Ieee. pp. 1398–1402.

Wild C.P., Weiderpass E., S.B.e., 2020. World cancer report: cancer research for cancer prevention. International Agency for Research on Cancer URL: https://publications.iarc.fr/586.

Yao, S., Tan, J., Chen, Y., Gu, Y., 2021. A weighted feature transfer gan for medical image synthesis. Machine Vision and Applications 32, 1–11.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 586–595.

Medical Imaging and Applications

Master Thesis, June 2023

Comparative Analysis and Explainability of Mono-input and Multi-input CNNs in Classifying Thyroid Nodules from 2D Ultrasound Images

José Carlos Reyes Hernández, Alain Lalande (PhD), Sarah Leclerc (PhD)

ICMUB laboratory, UMR CNRS 6302, University of Burgundy, Dijon, France

Abstract

Thyroid nodule classification is crucial for the accurate diagnosis and management of thyroid diseases. However, visual classification by physicians has limitations, including time consumption, potential inaccuracies, and the risk of subjecting patients to unnecessary and stressful Fine Needle Aspirations (FNAs). Moreover, inter- and intraobserver variabilities further complicate the process.

In this study, Deep Learning (DL) models are explored for thyroid nodule classification based on 2D ultrasound (US) images. The "Attention-Densenet121" network, developed in previous work, is optimized to improve its performance. The objective of reducing unnecessary FNA procedures is addressed. The Bethesda score was utilized as the groundtruth, and the nodules were classified as requiring FNA or not requiring FNA.

Two configurations of the network, namely the mono-input and multi-input settings, are investigated considering the two different orientations of US images of thyroid nodules. A carefully constructed dataset of US images is utilized for model training and evaluation. The models are assessed based on their accuracy, F1-score, sensitivity, and specificity. The model with the highest specificity, aligning with the primary objective, is considered the best. Additionally, different fusion strategies were employed with the multi-input model to explore the combined features from both image planes.

The results highlight the effectiveness of the mono-input models in accurately identifying samples from the negative class with high specificity. Conversely, the multi-input models demonstrate high sensitivity in correctly recognizing samples from the positive class. The trade-off between specificity and sensitivity is discussed, emphasizing the importance of achieving a balance for accurate classification.

To gain interpretability, heatmaps generated using different post-hoc explainability (XAI) algorithms are employed, shedding light on the models' understanding of nodule localization. Additionally, the analysis examines the impact of dataset characteristics and model metrics on performance.

In conclusion, this study introduces DL models for thyroid nodule classification with the primary objective of reducing unnecessary FNA procedures. The importance of dataset construction and model optimization is emphasized as crucial factors in achieving reliable and accurate results. The findings of this research contribute to the development of diagnostic models that can improve patient care by minimizing the need for unnecessary procedures and promoting more efficient and targeted medical interventions.

Keywords: Thyroid nodules, Deep Learning, Binary classification, Fusion strategies, Explainability, Bethesda score

1. Introduction

The thyroid gland, a vitally important organ situated at the base of the neck, has the primary function of regulating metabolism through the production of various hormones. A common condition that can affect this gland is the formation of thyroid nodules, which are abnormal growths within the gland. Although most of these nodules are benign, a small percentage, estimated to be between 5 % and 15 %, can be cancerous (Hayat et al., 2007). This potential for malignancy underscores the necessity for effective diagnostic tools to facilitate appropriate therapeutic management.

One such diagnostic tool is Fine Needle Aspira-

tion (FNA), a minimally invasive procedure that involves the use of a thin, hollow needle to extract cells from the thyroid nodule for microscopic examination. Despite its limitation of leading to some unnecessary procedures due to the prevalence of benign nodules, FNA remains a crucial diagnostic approach. It enables clinicians to estimate the likelihood of malignancy and subsequently plan suitable treatment strategies (Cooper, 2009).

Ultrasound (US) serves as the primary imaging modality for the examination of the thyroid gland, prized for its non-invasive nature, affordability, and provision of real-time imaging. However, US imaging is not without challenges, including the potential for inter-observer variability and a dependence on the skill level of the operator for both acquisition and interpretation of images (Hoang et al., 2018).

To address these challenges and standardize the ultrasound analysis of thyroid nodules, the Thyroid Imaging Reporting and Data System (TIRADS) was developed (Grant et al., 2015). This system assigns a score from 1 to 5 to nodules based on their ultrasound characteristics, with higher scores indicating a heightened risk of malignancy.

In parallel, the Bethesda System for Reporting Thyroid Cytopathology offers a standardized classification for FNA results, providing specific recommendations for each of its six categories (Cibas and Ali, 2009). However, despite these systematic advances, a significant number of benign nodules still undergo unnecessary FNAs, causing patient discomfort and increasing healthcare costs (Schnadig, 2018).

To address the issue of unnecessary FNAs, a previous internship led to the development of a Computer-Aided Diagnosis (CAD) system that utilizes Deep Convolutional Neural Networks (DNNs). This system sought to automate the prediction of unnecessary FNAs, using the Bethesda score as the ground truth, with the aim of improving the management of thyroid nodules and enhancing patient care.

For a medical diagnosis system to gain the trust of physicians, technicians, and patients, it must demonstrate transparency, comprehensibility, and explainability. Ideally, such a system should provide a clear rationale behind its decision-making process to all stakeholders involved (Singh et al., 2020). However, despite their effectiveness, the challenge with Deep Learning (DL) models often lies in the opacity of their inner workings. In particular, the weights of the neurons are not directly interpretable, which can hinder understanding and trust in these models (Meyes et al., 2020).

To bridge this gap and associate the outputs of the previously developed model with the fundamental descriptors used by clinicians for image interpretation and diagnosis, gradient-based explainable AI techniques such as Grad-CAM were employed to generate heat-maps (Selvaraju et al., 2017). These visual tools aimed to illustrate the model's decision-making process during diagnosis prediction. However, upon presenting these visual maps to physicians, feedback indicated that the activated regions in the resultant heatmaps did not correctly focus on the relevant areas of the nodule.

The main contributions of the following work include:

1. *The undertaking of a comparative analysis of mono-input and multi-input DL models:* This work provides a comprehensive evaluation of such models for thyroid nodule classification from US images, highlighting distinctive performance characteristics.
2. *The employment of explainable AI techniques for model decision interpretation:* Explainable AI techniques like Grad-CAM have been utilized to produce heatmaps, offering intuitive visualization of DL models' decision-making process.
3. *The comparison of different explainable AI technique-generated heatmaps:* A comparison of heatmaps generated through various explainable AI techniques has been conducted, uncovering the strengths and weaknesses of each method within this specific context.
4. *The thorough translation of the model's architecture from Tensorflow to PyTorch:* In alignment with the supervising institution's ongoing initiative to construct a proprietary DL models training library, the previously developed model's codebase was carefully translated from Tensorflow to PyTorch.

2. State of the art

The problem of thyroid nodule classification using DL algorithms techniques has drawn significant attention in recent years. The use of US imaging has proven highly effective in the detection of these nodules, and DL models, in particular, have shown promising results in automating their classification as benign or malignant. However, despite their impressive performance, the complexity and 'black box' nature of these models make it difficult to understand the rationale behind their predictions, leading to trust and adoption issues, especially in high-stakes domains like healthcare.

The explainability or interpretability, of DL models, is a key consideration being delved into. This refers to the ability to comprehend the inner workings of a model, and it is especially vital in medical applications where the impact of model predictions can directly shape clinical decisions and patient outcomes (Holzinger et al., 2019). Consequently, an increasing focus on the development of methods to enhance the interpretability of these models has been witnessed in the field.

Dataset	Description	FNA not required	FNA required
Baseline	Initial reference dataset.	242	213
Baseline chu	Contains only images from private sources.	242	124
Enlarged	Augmented version of the "Baseline" dataset.	417	335
Enlarged chu	Augmented version of the "Baseline chu" dataset.	417	246
Multi-view	Dataset specially curated for training the Multi-Input network.	328	170
Validation set	Dataset reserved for validation during training on the entire Baseline and Enlarged datasets.	24	20
Test set	Dataset designated specifically for inference purposes	20	20
Multi-view val set	Dataset reserved for validation during training on the entire Multi-view dataset.	36	28
Multi-view test set	Test dataset specifically utilized for inference purposes in the Multi-Input network.	36	28

Table 1: Summary of dataset subsets used in this study. Each subset was generated based on different criteria, such as the source of the images, and the need for an FNA procedure. The columns "FNA not required" and "FNA required" indicate the number of images in each subset that were assigned these labels based on the Bethesda score. The images belonging to the validation and test sets were chosen randomly while ensuring a balanced distribution of samples across the different classes.

In the following sections, the current state-of-the-art in DL models designed for thyroid nodule classification from US images, as well as the explainability techniques developed to enhance interpretability in this specific context, will be explored.

2.1. *Classification of malign nodules from 2D US thyroid images using DNNs*

Numerous studies have been conducted in the field of DL with regard to the classification of thyroid nodules using US images. Buda et al. (2019) proposed a DL algorithm that uses thyroid US images to provide management recommendations for thyroid nodules observed in US images. The algorithm was trained on a robust dataset comprising 1377 thyroid nodules from 1230 patients. The ground truth for each image was established based on the risk of malignancy, as determined by the TIRADS score. A multi-task DNN was trained to recommend biopsies for thyroid nodules, using two axial US images as inputs. The algorithm's sensitivity and specificity were compared with the consensus of three experts from the American College of Radiology (ACR) TIRADS committee, as well as nine other radiologists. For a test set of 99 nodules, the proposed system achieved a sensitivity of 87 %, on par with the expert consensus and surpassing five out of the nine radiologists. The algorithm's specificity was 52 %, similar to the expert consensus and exceeding that of seven out of the nine other radiologists.

Chi et al. (2017) introduced another DL-based system for categorizing thyroid nodules using US images, with the TIRADS score serving as the groundtruth. To ensure accuracy, the US images were initially processed to adjust their scale and eliminate any artifacts. A pre-existing GoogLeNet (Szegedy et al., 2015) model was then refined using these processed images, which enables enhanced feature extraction. These derived features were then inputted into a Cost-sensitive Random Forest classifier to categorize the images as either "malignant" or "benign". The experimental findings demonstrated that the refined GoogLeNet model delivered exceptional classification performance. It reached a classification accuracy of 98.29 %, a sensitivity of 99.10 %, and a specificity of 93.90 % when using an open-access database. Furthermore, it achieved a classification accuracy of 96.34 %, a sensitivity of 86 %, and a specificity of 99 % when using images from their local health region database.

In their research focusing on limited US thyroid image datasets, Zhu et al. 2017 highlighted the importance of implementing data augmentation techniques. They conducted a comparison between conventional methods and a specialized Convolutional Neural Network (CNN) designed for data augmentation. The process encompassed pre-processing steps to extract the region of interest (ROI), data augmentation via both standard procedures and a compact CNN, and classification of thyroid nodules into benign or malignant categories through transfer learning that leveraged a pre-trained residual network. This network was fine-tuned using three distinct datasets: the original, the traditionally augmented, and the CNN-augmented dataset. The results indicated that their approach, utilizing a CNN for data augmentation, yielded a superior accuracy rate of 93.75 %, comparable to other relevant methods.

2.2. *Towards explainability in DL for classification in thyroid nodule images*

To understand how a DL model makes predictions about whether a nodule is benign or malignant, it is essential to incorporate explainability techniques into the pipeline of a CAD system. In this regard, Yang et al. (2022) first trained a ResNet18 (He et al., 2016) model to predict the benign or malignant nature of a thyroid nodule. They then used a Grad-CAM algo-

rithm to generate heatmaps that would highlight sensitive regions in an ultrasound image during the learning process. This method enabled the extraction and analysis of shape features from these sensitive regions. The results revealed marked differences between benign and malignant thyroid nodules, indicating that the shape features of sensitive regions significantly assist in diagnosis.

In the study conducted by Kong et al. (2022), an Attribute-Aware Interpretation Learning (AAIL) model for thyroid ultrasound diagnosis was proposed. The AAIL model is composed of two modules: an Attribute Properties Discovery module and an Attribute-Global Feature Fusion module. The former module is designed to extract the key attributes of thyroid US images, while the latter is used to integrate these extracted attributes with the global features of the images. The AAIL model was trained on a dataset of thyroid US images along with their corresponding labels based on the TIRADS score. The experimental outcomes indicated that the AAIL model surpassed traditional machine learning models in terms of diagnostic performance. Furthermore, the AAIL model was able to provide interpretable results, aiding doctors in gaining a better understanding of the diagnostic process.

Similarly, Deng et al. (2022) introduced a multi-task attention network, guided by clinical knowledge, to analyze thyroid nodules. The network initially classifies each descriptor within the ACR TIRADS lexicon. These individual descriptor scores are subsequently combined to generate a comprehensive risk score for the nodule. This consolidated risk score is then utilized to categorize the nodule as either benign or malignant. The proposed methodology was assessed using a dataset comprising 1,000 US images, yielding an accuracy of 93.55 %, a sensitivity of 93.8 %, and a specificity of 93.14 %.

Likewise, Manh et al. (2022) presented the Multi-Attribute Attention Network (MAAN), as an innovative approach for the interpretative diagnosis of thyroid nodules within US images. MAAN incorporates an attention mechanism enabling the model to discern the significance of various image features in diagnostic evaluations. This attention mechanism allows MAAN to concentrate on the most informative aspects within each image, thereby enhancing diagnostic precision while simultaneously offering insights into the foundational causes of the nodules. MAAN underwent evaluation using a dataset comprising 1,000 ultrasound images of thyroid nodules, yielding an impressive accuracy rate of 95.2 %, on par with the precision of human experts.

The present work differentiates itself from previous studies by incorporating the Bethesda score as the ground truth for thyroid nodule classification, which aligns with the objective of reducing unnecessary FNA procedures. Furthermore, in addressing the limitations of the TIRADS in identifying relevant image features, this study emphasizes the need for more insightful post hoc explainability methods to gain deeper insights into the diagnostic process.

3. Material and methods

3.1. Dataset

A comprehensive summary of the dataset is provided in Table 1. This table aims to provide a clear overview of the composition and scope of each subset within the dataset.

The global dataset utilized in this study consists of various subsets of thyroid nodule images, all of which are extracted in JPEG (Joint Photographic Experts Group) format. In this study, the images are saved in JPEG format as they are screen captures from US scanners. The total number of US thyroid nodules in the dataset is 747. These images were sourced from two different hospitals in France: the Hospital of Dijon and the Hospital of Bastia. Two distinct scanners were used in the acquisition of these images, specifically the Aixplorer and Canon. Consequently, the image sizes differ based on the scanner used, with dimensions of 1440 x 1080 pixels and 1260 x 960 pixels for the Aixplorer and Canon scanners, respectively. Also, each thyroid nodule is typically represented in two orientations, with both axial and sagittal views acquired per case. This approach ensures a thorough evaluation of each nodule.

Importantly, all images were annotated by experts, and, following a visual examination based on the TIRADS criteria, an FNA procedure was conducted for each case involved in this study. This examination led to the categorization of the images into two labels: ”FNA not required” and ”FNA required”, as determined by the Bethesda score referenced in Section 1. The former label was assigned to 461 images, and the latter to 286 images.

In an effort to ensure a more balanced dataset, an additional 89 publicly available thyroid nodule images, which were annotated and identified as malignant, were incorporated (Dasmehdi and Xtr, 2017). This decision to use this public dataset was made despite the potential for adding noise to the model's training process, given that the images from this dataset have a lower image size of 348 x 272 pixels.

3.1.1. Subsets

In the process of creating these distinct subsets, meticulous care was taken to separate the images based on the assigned Bethesda score. This score, determined through expert annotation, results from a cytopathological examination of thyroid nodules.

Additionally, each image's orientation, axial or sagittal, was annotated manually. Given that each thyroid nodule is typically represented in both orientations, this distinction was imperative for a comprehen-

sive evaluation of each nodule. This careful annotation process ensured that the subsets were accurately formed, reflecting the specific characteristics of each image, such as its Bethesda score and orientation.

This process allowed for a systematic classification and separation of images, thereby facilitating the creation of well-defined subsets tailored to the specific requirements of the training pipeline.

The initial experiments utilized a dataset similar to the one used in the preceding internship, herein referred to as the ”Baseline” dataset. Since then, new images have been collected, resulting in an expanded dataset that combines the baseline dataset with these newly acquired images, forming the ”Enlarged” dataset. It is worth noting that the most recent images were sourced exclusively from the Hospital of Bastia, where the Canon scanner was used to acquire these images.

On the other hand, as the focus shifts to considering only images sourced privately, two new subsets were created by excluding images from the public dataset as referenced in Section 3.1. These two subsets are derivatives of the ”Baseline” and ”Enlarged” datasets.

Furthermore, a dedicated subset, named the ”Multi-view” dataset, was built for the purpose of training a multi-input network. This subset comprises solely pairs of images, each consisting of an axial and its corresponding sagittal view. Accompanying this subset are its respective ”validation” and ”test” sets.

3.2. Pre-processing

Given the difference between the US scanners from which the images were collected, as referenced in Section 3.1, some image enhancement techniques were deemed necessary to ensure consistency and uniformity in size and exposure, as well as to eliminate artifacts present in the images that might hamper the network's learning process. An example of a raw image from the provided dataset is shown in figure 1.

The pre-processing pipeline is outlined below:

Cropping: As an initial step, the US images were systematically cropped to remove unnecessary information, leaving only the data produced by the US waves. These images, as mentioned in Section 3.1, are screen captures and initially presented in a three-channel format. They were first converted into grayscale. Then, by sequentially scanning the rows and columns of each grayscale image, boundaries were identified where the average pixel intensity exceeded a predetermined threshold. This threshold signified the beginning of significant visual content. The rightmost limit within the threshold was also identified to establish the image's boundaries. The images were subsequently cropped based on these determined limits, effectively eliminating the black borders and ensuring that only relevant information was retained.

Removing Artifacts: The images in the dataset contained a scale bar that needed to be removed in order to focus exclusively on the thyroid information. Following the cropping operation, a binary image was subsequently generated from the grayscale image using a thresholding operation. From this binary image, all object contours or shapes were identified, with the longest one, presumed to be the scale bar, selected for elimination. A binary mask was then created using this contour and slightly expanded to ensure the entire scale bar was covered. The region within the mask was filled with the texture from the surrounding area, effectively erasing the scale bar from the image. Finally, the grayscale image was converted back into a three-channel image without the scale bar, ensuring that the subsequent analysis focused solely on the relevant US data.

The resulting preprocessed image can be seen in Figure 2.

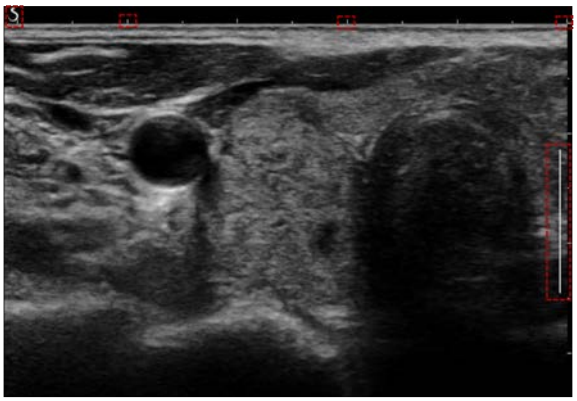


Figure 1: Raw US image of the thyroid nodule. The dashed rectangles serve to indicate the locations of the artifacts. Additionally, black borders can also be observed.

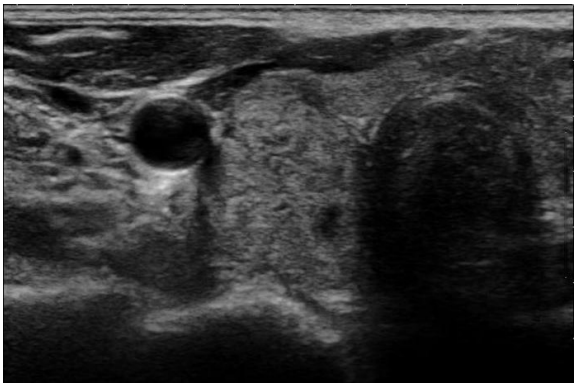


Figure 2: The resulting preprocessed US image is free from artifacts and retains only the relevant information.

3.3. Data Augmentation

The importance of Data Augmentation (DA) in DL, particularly when dealing with a small dataset, cannot be overlooked. With limited data, a model is at risk

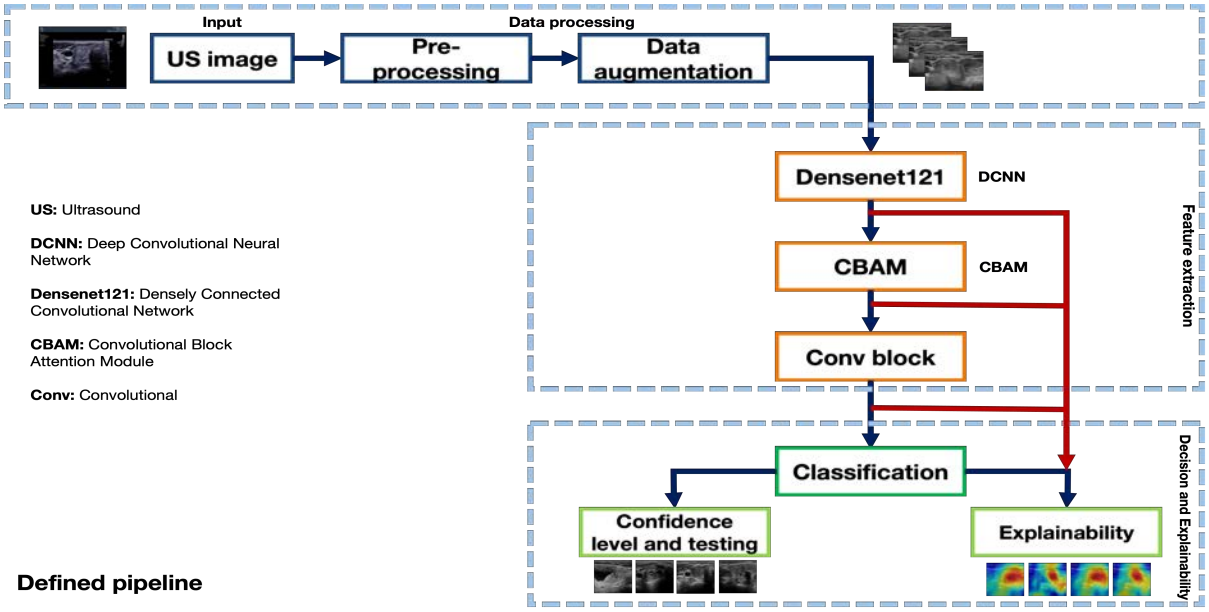


Figure 3: An overview of the established pipeline is presented. Data augmentation was applied exclusively to the training set. Additionally, the red arrows denote the chosen target layers for generating heatmaps using CAM algorithms.

of overfitting, where it learns the training data so well that it struggles to generalize to new, unseen data. To mitigate this, DA transformations are applied to the US images used in this study, enhancing the robustness of the DNN model.

The network used in this study was developed in PyTorch, leading to all image augmentations and transformations being performed using the 'Torchvision' library. It is important to note that a combination of on-the-fly DA and preprocessing operations is standard practice in PyTorch.

Given the characteristics of the dataset, the following DA transformations and preprocessing steps have been considered:

- **'Resize'** is applied to all images to ensure uniformity in size before they are fed into the network. A standard size of 800x600 pixels is used for all input images. Bilinear interpolation is used during the resizing process, aiding in the preservation of as much of the original image detail as possible.
- **'RandomHorizontalFlip'** introduces variability into the dataset by simulating possible changes in orientation that might occur during US scans.
- **'RandomRotation'** is applied to the images to enhance the model's ability to handle both minor shifts in the nodule's position and various orientations of the entire image. This transformation simulates the variability that might occur in real-world US imaging, where images can be taken from different positions.
- **'ColorJitter'**, which randomly adjusts brightness and contrast, enables the model to recognize nodules under varying contrast conditions.
- **'RandomAffine'**, including random translations and scaling, simulates variations in the position and size of the nodules.
- **'ToTensor'** is a preprocessing step that converts the images into PyTorch tensors, a format that is necessary for PyTorch models.
- **'Normalize'** standardizes the pixel values of the images using predefined statistics. The mean and standard deviation values used for normalization were computed directly from the dataset.

Both the DA transformations and preprocessing operations are applied on-the-fly to each batch during every iteration of the model's training process.

3.4. Proposed pipeline

The pipeline for this project, established during a previous internship, consists of several key steps. These include pre-processing, DA, and feature extraction utilizing a pre-trained Densenet121 as the main backbone. Additional components were incorporated, such as Convolutional Blocks Attention Modules (CBAM), and a final convolutional block. The latter is comprised of a Separable Convolution, Batch Normalization, a Rectified Linear Unit (ReLU) activation function, and an Adaptive Max-pool layer. A Fully-Connected (FC) layer was also included for

binary classification purposes. This network configuration was subsequently denoted as the 'Attention-Densenet121'.

Initially, due to the fact that the proposed architecture had been previously coded in Tensorflow, it was necessary for the entire codebase to be translated into the PyTorch framework. This transition was motivated by an ongoing initiative to construct a proprietary library for training DL models within the institution that supervised this project. It should be noted that during the code translation process, careful attention was given to ensure accurate replication of all model configurations, including non-linearities, weight initialization, and the number of parameters. Once the full pipeline was successfully migrated into the target framework, efforts were made to optimize the hyperparameters, and training strategy, and to select the most suitable dataset for training. These measures aimed to enhance the network's performance.

Subsequently, in pursuit of more profound insights into the network's rationale, a variety of post-hoc explainability algorithms were tested for heatmap generation, providing visual interpretations of the model's decision-making process.

On the other hand, in a bid to further enhance the model's predictive capabilities, two distinct network configurations were explored: a Mono-Input Network and a Multi-Input Network. These configurations were designed to use thyroid nodule images in different manners, thereby exploiting the unique features and information each view of the thyroid nodules provides.

Further details regarding the implemented procedures will be presented in the following sections. A comprehensive illustration of the pre-defined pipeline can be found in Figure 3.

3.4.1. Hyper-parameters optimization

To enhance network performance, an initial experiment was conducted using the baseline dataset as detailed in Section 3.1.1, with the model being trained under the original set of hyperparameters. This was followed by a Random Search, a process of hyperparameter optimization that involves the selection of random combinations from a predetermined range of hyperparameters to identify the best solution. During this phase, several key objectives were pursued. These included the optimization of the batch size, learning rate, learning rate scheduler, dropout, and the selection of an optimizer. Additional techniques incorporated to boost network effectiveness included weight decay and gradient clipping. Weight decay refers to a regularization method that prevents overfitting by adding a penalty term to the loss function, thereby reducing the magnitude of the weights (Loshchilov and Hutter, 2017). Gradient clipping, on the other hand, is a technique to prevent exploding gradients by limiting the maximum value of gradients (Zhang et al., 2019).

Moreover, in this context, a stratified five-fold cross-validation approach was employed. This was done to more accurately estimate the model's capacity to generalize to unseen data, given the selected sets of hyperparameters.

3.4.2. Backbone's weights

The dataset's unique image characteristics encouraged experiments assessing the impact of applying pre-trained weights to the Densenet121 backbone, compared to training the network from scratch. This was conducted to evaluate transfer learning benefits, where the model is initialized with weights learned from a different, typically larger, dataset - in this case, ImageNet. It is important to note that these weights are not updated during backpropagation. The relevance of this approach arises from the stark differences between the ImageNet images and the US thyroid nodule images. The goal was to establish the effectiveness of transfer learning in this specific context, determining if features learned from the large-scale dataset could be efficiently applied to thyroid nodule classification.

Additionally, experiments were conducted using different dataset versions. This was driven by the hypothesis that certain dataset characteristics and size could impact the decision to use pre-trained weights, potentially influencing the experiment outcomes.

3.4.3. Optimization of hidden layers in the classifier

To boost the network's performance, the structure of the classifier's hidden layers was analyzed. Experiments were conducted to reveal how changes in the number of hidden layers affected the model's output.

The foundation for these experiments lies in the fundamental architecture of DL models. The depth of a model, indicated by its hidden layers, can significantly impact its capacity to discern complex patterns and relationships within the data. Models with a higher count of hidden layers are known for their proficiency in comprehending hierarchical feature representations, which could be highly beneficial given the complex nature of the application.

Nonetheless, a balance must be struck. An increase in the number of hidden layers can potentially enhance the model's learning ability, but it may also lead to overfitting, especially when working with a limited-sized dataset.

Bearing these considerations in mind, a systematic adjustment of the hidden layers in the classifier was undertaken. The objective was to pinpoint an optimal balance between the model's learning ability and the risk of overfitting, with particular emphasis on the dataset's unique characteristics.

3.4.4. Mono-Input network

In this context, it is important to note that each thyroid nodule image was fed into the network individu-

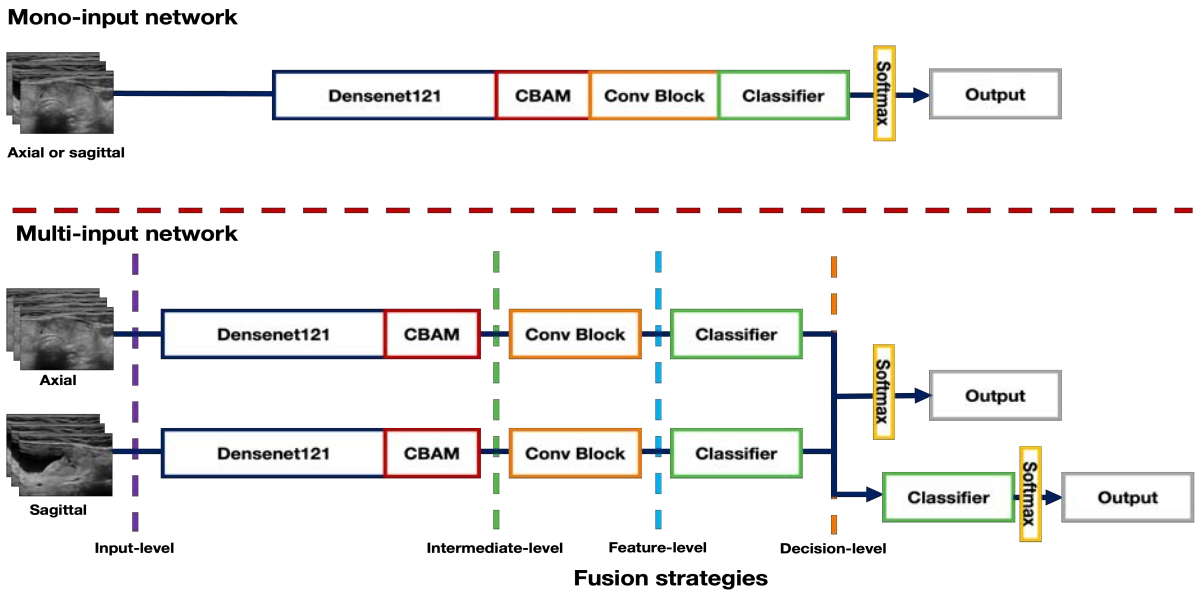


Figure 4: A schematic overview of the networks and fusion strategies used in the training process is depicted. The top part of the figure presents the Mono-Input network, while the bottom part depicts the Multi-Input network. The layers where the feature maps are concatenated for each of the fusion strategies are indicated by colored dashed lines: purple for input-level fusion, green for intermediate-level fusion, light blue for feature-level fusion, and orange for decision-level fusion, respectively.

ally, either in an axial or sagittal orientation. This approach was taken to ensure that each orientation was processed independently, thereby allowing the model to capture unique features present in each view. By handling the images separately, the network can effectively learn and identify the distinct characteristics of nodules in both axial and sagittal planes. This approach increases the model's ability to discriminate between images requiring FNA and those that do not, enhancing the overall accuracy and robustness of the predictions.

However, it should be acknowledged that this mono-input approach can have potential drawbacks. One significant issue is that it may result in different classifications for the same nodule, depending on the view presented to the model. This inconsistency poses a challenge, as it leaves open the question of which classification to follow when there is a discrepancy between the results obtained from different orientations.

3.4.5. Multi-Input network

The decision to simultaneously input both axial and sagittal images into the network is a result of the distinct perspectives they offer for the evaluation of thyroid nodules. These two views reveal different features of the nodules when examining the thyroid, aiding in the detection of any anomalies. In this setup, each image type follows a separate branch within the network, ensuring that their unique information is independently processed. This combined approach provides the network with a more comprehensive understanding of the nodule's structure, potentially leading to more accurate predictions regarding the necessity of an FNA.

In this approach, a prediction is made by the model for each pair of images that are inputted. Accordingly, the dataset that was used in this context was built to ensure that an axial and a sagittal view of the image is always present for each exam.

Moreover, to further enhance the model's ability to distinguish complex patterns and nuances in the data from these two distinct anatomical perspectives, various fusion strategies were employed. These strategies were designed to allow the model to handle multi-view data effectively, thereby exploiting the full potential of the axial and sagittal images.

3.4.6. Fusion strategies

The adoption of different fusion strategies— input-level, feature-level, intermediate, and decision-level —facilitates the integration of learned features from the axial and sagittal views in distinct ways, each with its own advantages.

Input-level fusion involves the concatenation of axial and sagittal images into a single tensor prior to their introduction into the network. This strategy allows the network to concurrently process both views from the very beginning. As a potential benefit, this approach could enable the network to learn features that are intrinsically interdependent on both views right from the start. This may be especially beneficial when the features in the axial and sagittal views have strong correlations or interact in ways that sug-

Method	Authors	What it does	Pro	Con
GradCAM	Selvaraju et al. (2017)	Use of gradient information to highlight important regions for predictions	Simple and effective	Lower resolution heatmap
GradCAM++	Chattopadhay et al. (2018)	An extension of GradCAM that captures multi-level and multi-object features	Better at handling multiple objects	More computationally expensive
XGradCAM	Fu et al. (2020)	An improved version of GradCAM with better visual fidelity	Provides more detailed and accurate visualization	More computationally intensive
AblationCAM	Ramaswamy et al. (2020)	Ablates each activation map and measures the decrease in output score	Can handle model architectures that GradCAM cannot	Very computationally intensive
ScoreCAM	Wang et al. (2020)	Generates heatmaps by activating neurons in the target layer one by one	Does not require gradients, making it applicable to a wider range of models	Very computationally intensive and slow
EigenCAM	Muhammad and Yeasin (2020)	Applies PCA to the feature maps and uses the principal components to highlight regions	Can capture more diverse features	Might be harder to interpret

Table 2: Brief comparison of heatmap generation methods used in DL. Each method is assessed based on its functionality, advantages, and limitations. The methods vary in their computational complexity, the nature of the information they use to generate heatmaps (e.g., gradients, activation maps, or principal components), and their ability to handle different model architectures or scenarios (e.g., multiple objects, diverse features).

gest thyroid nodule anomalies (Seeland and Mäder, 2021).

Moving on to **feature-level fusion**, the features from both views are combined immediately after the last feature extraction layer. This approach enhances the model's ability to learn shared representations from the very beginning of the classification stage. The benefit here is that the model can capture interactions between the two views at all succeeding levels of representation, potentially leading to a richer, more complex feature space that takes into account the relationship between axial and sagittal images early on (Seeland and Mäder, 2021).

Intermediate-level fusion, in contrast, processes the two inputs separately up to the first spatial attention module. The features are then combined and processed together through the rest of the network. This strategy is beneficial when there are important view-specific features that should be learned separately before they are combined. It allows the model to learn and maintain view-specific representations, and then merge these representations to capture the relationships between the views (Zhang et al., 2021).

Lastly, the **decision-level fusion** strategy processes each input through the entire network, including the initial classifier, separately. The preliminary predictions are then combined and passed through a final classifier. This approach is advantageous when the axial and sagittal views contain largely independent information that can contribute to the final decision. By allowing each view to reach a preliminary decision independently, late fusion can exploit the unique information present in each view to the fullest extent (Seeland and Mäder, 2021).

A schematic overview of the networks and fusion strategies previously described is presented in Figure 4.

3.4.7. Post-hoc explainability

Post-hoc explainability (XAI) in DL refers to the process of interpreting the decisions of a model after its training phase. As highlighted in Section 2, the complexity and lack of transparency in DL models often lead to them being referred to as "black boxes". The aim of post-hoc explainability is to shed light on these black boxes, thus revealing what the model has learned and the mechanisms behind its decision-making process. A variety of techniques are utilized to accomplish this, such as visualizing the network's activations and weights, and employing saliency maps to emphasize significant features in the input data (Holzinger et al., 2022a).

The proposed method in this study emphasizes the generation of heatmaps using attention modules and Class Activation Maps (CAM) algorithms. Both these tools yield heatmaps, which are visual representations indicating where a model is "paying attention" during prediction. However, the means of generating these maps and the precise insights they offer can vary.

It is crucial to distinguish between attention maps and techniques like GradCAM. Although both provide visual indications of a model's focus, they do so in fundamentally different ways.

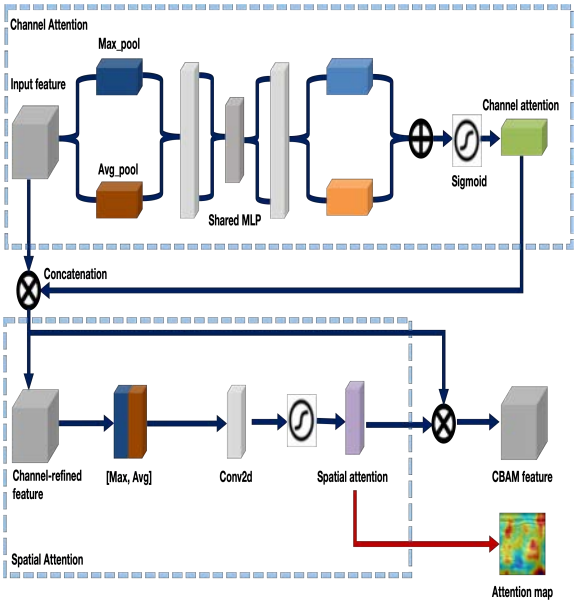


Figure 5: A schematic representation of the channel and spatial attention mechanisms. The upper section illustrates the channel attention module, while the lower section depicts the spatial attention module. The red arrow indicates the layer where the attention map is generated.

3.4.8. Attention maps

Attention modules in a network go beyond visualization. They actively guide the model's focus by promoting the most meaningful regions and diminishing the less significant information. This process ultimately shapes the model's attention, offering a more directed approach to identifying key features within the data. (Guo et al., 2022).

As seen in Figure 3, the attention mechanism included in the network was the CBAM introduced by Woo et al. (2018). CBAM operates by sequentially employing channel and spatial attention modules to adaptively adjust feature maps across both dimensions. The importance of each feature map is assessed by the channel attention module, while the spatial attention module further refines the feature maps by taking into account the inter-spatial relationships among features. A schematic overview of this mechanism is presented in Figure 5.

In this work, the attention maps were generated directly from the attention weights in the Spatial Attention module.

3.4.9. CAM algorithms

CAM algorithms, in general, work by providing a visual representation of the importance of different regions in the image in relation to the model's final decision. This is achieved by using the gradient information flowing into the model's convolutional layers. By highlighting the areas that significantly impact the prediction, these algorithms offer a comprehensive view of the model's decision-making process (Holzinger et al., 2022b). The advantage of these algorithms lies in their ability to take into account the entire network, instead of focusing solely on a specific layer. This allows for a more complete understanding of how the model interprets the input data to arrive at its final decision. A summary of these algorithms can be found in Table 2.

As depicted in Figure 3, three distinct layers were selected to generate the heatmaps, based on the architecture of the network used in this work. These layers were chosen because they are likely to contain the most semantically meaningful (i.e., informative) activation maps. The selected layers are as follows:

- **The final layer of the backbone**: This layer directly follows the backbone with pre-trained weights and may contain high-level feature maps beneficial for diagnosis.
- **The spatial attention layer**: This layer follows the spatial attention mechanism and may contain feature maps further refined by attention.
- **The final convolutional layer**: This layer is the last convolutional layer before the classifier, and the output from this layer directly feeds into the classifier. This layer is likely to possess the most abstract and high-level feature representations that are most closely tied to the final class predictions.

All these heatmaps were generated using the PyTorch Grad-CAM library, a package created by Gildenblat and contributors (2021) that includes state-of-the-art methods for XAI in computer vision.

4. Model training

The training was monitored and convergence optimized using Rectified Adaptive Moment Estimation (RADAM) with a learning rate set to 0.001. Unlike ADAM, RADAM rectifies the variance of the adaptive learning rate to provide a more consistent and reliable convergence, making it a superior choice for this optimization (Liu et al., 2019).

Given the imbalanced nature of the provided dataset, Focal Loss was employed as the loss function (Lin et al., 2017). This decision was made in order to address the imbalance by automatically computing the loss weights based on the data distribution, thereby assigning more weight to the underrepresented class. The gamma parameter was fixed at 3.5.

Gradient clipping, which limits the magnitude of gradients to prevent exploding gradients and improve model stability, was found to be advantageous and was implemented with a clip value of 5. To aid in regularization, weights decay was introduced with a value set to 0.001, which helps in preventing overfitting by adding a penalty to the loss function based on the size of the weights.

To adjust the learning rate during training, a type of learning rate scheduler known as "ReduceLROnPlateau" was utilized. This is a specific method that dynamically adapts the learning rate based on the model's performance. Specifically, this scheduler diminishes the learning rate when the model's performance ceases to improve, thereby providing a more effective learning rate for the training process. It is characterized by a 'patience' parameter, which was set to 5 in this case, dictating the number of epochs with no improvement after which the learning rate will be reduced (Biskup, 2008).

Additional measures to prevent overfitting included early stopping and dropout. Early stopping was implemented with a patience value of 40, and the model was selected where the validation loss was at its lowest. Dropout was set to 0.25, a technique that randomly deactivates neurons in the classifier to improve generalization by preventing the model from becoming overly reliant on any single neuron.

The batch size for the training was fixed at 16, and the models were trained over 150 epochs. Weights that were not pre-trained in the architecture were initialized using the Kaiming distribution. This method of initialization is beneficial because it helps to prevent the issue of vanishing and exploding gradients during backpropagation. By maintaining the variance of the weights, the Kaiming initialization ensures that each neuron operates in a region where it is sensitive to the inputs and can learn from them effectively. This leads to improved training speed and performance (He et al., 2015).

To ensure the reproducibility of the model training process and the subsequent results, a fixed random seed was established for any source of randomness. This practice ensures that the randomness in any part of the process is consistent, enabling reliable replication of the training process.

4.1. Stratified k-fold cross-validation

Stratified 5-fold cross-validation was employed during the training phase as a dual-purpose strategy to enhance model performance. This method of cross-validation is particularly useful when dealing with imbalanced datasets, as it ensures that each fold contains a proportional representation of each class. This helps in retaining the distribution of the classes and mitigating the impact of class imbalance during model training and validation (Szeghalmy and Fazekas, 2023).

The first purpose of using stratified 5-fold cross-validation was to explore the generalization capabilities of each model. By dividing the data into five distinct subsets, or "folds," derived from the respective training set of each version of the dataset, and iteratively training and validating on different combinations of these folds, a more comprehensive understanding of each model's ability to generalize beyond the training data could be gained.

In the process of performing the cross-validation, meticulous care was taken to apply DA exclusively to the training sets in each fold. This was essential to ensure that the validation set remained untouched and accurately represented unseen data, thus providing a reliable evaluation of the model's performance. This approach further ensured that the model was not inadvertently exposed to augmented versions of the validation data during training, which could potentially bias the validation results.

The second use of stratified 5-fold cross-validation was to construct an ensemble of models, with an emphasis on leveraging the best-performing models from each fold to enhance the robustness of their predictions. The best-performing models were determined by their lowest validation loss value achieved during training, and these models were saved for the subsequent ensemble construction. These models were combined using the stacking ensemble method, which involves stacking the predictions of each model to create a new input representation. This stacked input is then fed into a classifier to generate the final predictions. By combining the predictions of multiple models, the stacking ensemble improves the accuracy and robustness of the overall predictions. Ensemble models, such as the stacking ensemble, are known for their ability to mitigate the individual weaknesses of constituent models and deliver improved performance

5. Results

This section concisely presents the results of the conducted experiments.

5.1. Baseline model optimization

Initial experiments were conducted to improve the performance of the Attention-Densenet121 architecture. The model was initially trained on the "Baseline" dataset using the original set of hyperparameters and later extended to the "Enlarged" dataset. Stratified 5-fold cross-validation, as described in Section 4.1, was employed to assess the model's generalization capabilities comprehensively. The primary metric optimized during training was the validation loss, aiming to establish a highly confident model in its predictions. Although accuracy was not the main metric of focus, the mean values of this metric are presented as a reference to evaluate the model's performance before inferring on the test set. It is worth noting that the use of stratified 5-fold cross-validation ensured balanced partitions within each fold.

Numerous experiments with varying hyperparameters were conducted initially. After testing multiple configurations, the use of Random Search was explored. This method proved to be effective in identifying sets of hyperparameters that improved model performance, as evidenced in Table 3. The model configuration identified through this method was then employed in subsequent experiments.

Model	Validation loss	Validation accuracy
Baseline model	0.0518	0.8211
Optimized model	**0.0426**	0.8365

Table 3: Summary of the 5-fold cross-validation results: The baseline model refers to the model trained using the predefined hyperparameters, while the optimized model refers to the model trained using the hyperparameters obtained after hyperparameter optimization. Best result is shown in bold.

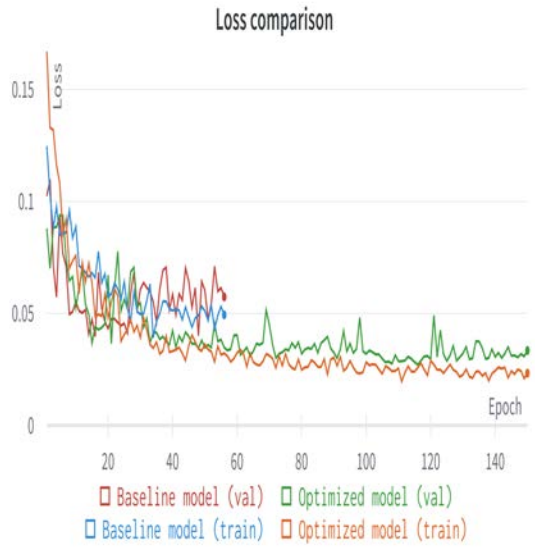


Figure 6: Comparison of loss curves for the baseline model and the optimized model configuration during a single fold of the cross-validation process. The newly identified set of hyperparameters facilitated the model's convergence and enabled it to reach a lower loss value.

5.2. *Benefits of pre-trained weights*

As discussed in previous sections, the unique nature of the provided dataset—relative to the datasets typically used to train pre-trained models in PyTorch—motivated experiments to train the network from scratch. These experiments were performed using the new, more effective set of hyperparameters and applied to different dataset subsets to investigate the influence of dataset size on network performance. Table 4 reveals a significant performance difference when using pre-trained weights versus not using them. Additionally, it demonstrates a decline in network performance when trained on smaller subsets, specifically referring to the datasets that only contain images from private sources. Counterintuitively when the network was trained on the "Enlarged" dataset, the performance also decreased.

5.3. *Evaluating the influence of hidden layers in the classifier*

After assessing the influence of using pre-trained weights, the next set of experiments aimed to investigate whether increasing the classifier's complexity could enhance the network's predictive abilities. These experiments primarily involved models trained on the "Baseline" and "Enlarged" datasets, since previous findings suggested that a smaller training dataset could potentially diminish the network's performance. While initial observations suggested that the inclusion of additional hidden layers might negatively affect the network's performance, the addition of an extra layer actually led to an overall improvement, as demonstrated in Table 5. These experiments facilitated the identification of the optimal network configuration and hyperparameter settings. The model deemed to be the best among those tested was subsequently utilized in further experiments involving its multi-input variant.

5.4. *Performance of Multi-Input network and Fusion strategies*

Upon identifying an optimal mono-input network configuration and set of hyperparameters, the experiments shifted focus to assess the performance of this network in a multi-input setting. This involved exploring the fusion strategies detailed in Section 3.4.6 to leverage the image features across different planes. It is important to note, as detailed in Table 1, that training and validation of this model required the use of a specific subset of the dataset, grouping the images into pairs, both in the training and validation sets. As shown in Table 6, the performance of this network configuration is less robust than its mono-input variant. However, within this context, the decision-level fusion strategy demonstrated better performance.

Fusion strategy	Validation loss	Validation accuracy
Input-level	0.1049	0.5864
Feature-level	0.1267	0.6128
Intermediate-level	0.3756	0.5623
Decision-level	**0.1033**	0.5594

Table 6: Comparison of the 5-fold cross-validation results of the multi-input network using different fusion strategies. This model was trained on a specific data subset for multi-input models, referred to as the "multi-view dataset." The final architecture of the model and the hyperparameters used for training in this comparison were chosen based on the "best model" in the mono-input setting. Best result is shown in bold.

5.5. *Binary classification results*

Upon completion of the various experimental sets, both the mono-input and multi-input networks underwent rigorous training. The mono-input network was trained on the "Baseline" and "Enlarged" datasets, while the multi-input network was trained on the "Multi-view" dataset. During the training process, each network variant was validated using its corresponding validation set.

The performance of these models is presented in Table 7. This table demonstrates the accuracy, F1-score, sensitivity, and specificity of six different models labeled A through F. Given that the primary objec-

Dataset	Pre-trained backbone		Non pre-trained backbone	
	Validation loss	Validation accuracy	Validation loss	Validation accuracy
Baseline	**0.0426**	0.8365	0.0656	0.6442
Baseline chu	0.0758	0.7441	0.0779	0.5330
Enlarged	0.0599	0.7050	0.0783	0.5429
Enlarged chu	0.0822	0.6258	0.0744	0.4905

Table 4: Comparison of the 5-fold cross-validation results of the optimized model across different data subsets, considering the use of pre-trained weights in the backbone network or not. Best result is shown in bold.

Dataset	n=2		n=3		n=4	
	Validation loss	Validation accuracy	Validation loss	Validation accuracy	Validation loss	Validation accuracy
Baseline	0.0426	0.8365	0.0455	0.8234	**0.0412**	0.8344
Enlarged	0.0599	0.7050	0.0617	0.6981	0.0552	0.7101

Table 5: Comparison of validation loss and accuracy across different sizes of hidden layers in the classifier for the optimized mono-input network. Results were obtained through 5-fold cross-validation on different data subsets. The "best model" configuration was selected based on these experiments, with the best result highlighted in bold. Here, "n" refers to the number of hidden layers in the classifier.

tive of this study is to reduce unnecessary FNA procedures, the model with the highest specificity is considered the best. Each model represents a particular configuration of the mono-input network:

- Model A represents the model that achieved the lowest validation loss across the five different folds during cross-validation on the Baseline dataset.
- Model B is the model that was trained on the entire Baseline dataset.
- Model C is an ensemble model that incorporates the weights of the models that achieved the lowest validation loss in each of the folds during cross-validation on the Baseline dataset.
- Model D represents the model that achieved the lowest validation loss across the five different folds during cross-validation on the Enlarged dataset.
- Model E is the model that was trained on the entire Enlarged dataset.
- Model F is an ensemble model that incorporates the weights of the models that achieved the lowest validation loss in each of the folds during cross-validation on the Enlarged dataset.

These results were presented in this manner to underscore the significant role that class distribution plays in the training process of the model. Both Model A and Model D, despite being trained on a subset of their respective datasets, demonstrated the highest performance across all metrics. For a more intuitive understanding of each mono-input model's performance, readers are referred to Figure 7.

Model	Accuracy	F1-score	Sensitivity	Specificity
A	0.75	0.73	0.67	**1.00**
B	0.57	0.52	0.54	0.71
C	0.68	0.67	0.64	0.73
D	0.75	0.74	0.69	0.86
E	0.73	0.72	0.68	0.80
F	0.73	0.72	0.67	0.84

Table 7: Classification results obtained using various configurations of the mono-input network. **A**: The model that achieved the lowest validation loss across the five folds during cross-validation on the Baseline dataset. **B**: The model trained on the entire Baseline dataset. **C**: An ensemble model that uses the weights of each of the models with the lowest validation loss from the five folds during cross-validation on the Baseline dataset. **D**: The model that achieved the lowest validation loss across the five folds during cross-validation on the Enlarged dataset. **E**: The model trained on the entire Enlarged dataset. **F**: An ensemble model that uses the weights of each of the models with the lowest validation loss from the five folds during cross-validation on the Enlarged dataset. The inference was performed on the test set. Best result is highlighted in bold.

Model	Accuracy	F1-score	Sensitivity	Specificity
A	0.47	0.36	1.00	0.45
B	0.47	0.36	1.00	0.45
C	0.53	0.47	1.00	0.48
D	0.56	0.56	0.63	**0.50**

Table 8: Classification results of the multi-input network using the selected best configuration of the mono-input network and employing different fusion strategies. **A** refers to the multi-input model utilizing input-level fusion, **B** refers to the multi-input model employing feature-level fusion, **C** refers to the multi-input model utilizing intermediate-level fusion, and **D** refers to the multi-input model employing decision-level fusion. The test set used for inference is derived from the "Multi-view dataset". Best result is shown in bold.

When examining the performance of the multi-input network on the same binary classification task, it is clear from Table 8 that the multi-input model using decision-level fusion achieved the best overall metrics, as previously suggested by Table 6. Notably, all these models demonstrated high sensitivity. For a clearer understanding of the performance of each multi-input model, please refer to Figure 8.

5.6. Results from post-hoc XAI techniques

In line with the study's aim to examine various XAI techniques for better comprehension of the decision-making process of the model, this section presents a series of heatmaps generated from different methods. As detailed in Section 3.4.7, these heatmaps were produced using various CAM algorithms and by using the weights extracted directly from the spatial attention module. They were also derived from different layers within the network employed for binary classification. It is worth noting that the model from which these heatmaps were generated is the one that demonstrated the best overall performance in the binary classification task, as described in the preceding section 5.5.

5.6.1. Backbone's heatmaps

From the qualitative results displayed in Figure 9, it becomes apparent that meaningful information regarding the nodule's localization is not revealed by any of the various maps generated for each case across the different algorithms, when the heatmaps are created using the gradient information up to the last layer of the backbone. However, for the instance where the network has accurately predicted the input image as 'FNA-not-required', the GradCAM begins to highlight the correct localization of the nodule. In contrast, for the 'FNA-required' case, a larger activation area is displayed by the EigenCAM, but it completely misses the nodule's localization.

5.6.2. Spatial module's heatmaps

As per the qualitative results depicted in Figure 10, a noticeable contribution by the spatial attention module within the CBAM is observed. For the input image correctly identified as 'FNA-not-required', the nodule's localization is either fully or partially highlighted by all the different algorithms, with the GradCAM and XGradCAM covering the largest part of the nodule. The only exceptions are the ScoreCAM and EigenCAM, which extend the highly activated area to structures not corresponding to the nodule. On the other hand, for the input image correctly classified as 'FNA-required', despite the model's high confidence in its prediction, none of the algorithms displays the correct localization of the nodule. Counterintuitively, in the case where the model misclassified the input image as 'FNA-required', the most activated areas are at least close to the nodule's position.

5.6.3. Last convolutional layer's heatmaps

The information up to this layer has a strong connection with the model's final decision. As shown in the qualitative results in Figure 11, a striking similarity across the different CAM methods and various cases is seen in all the heatmaps. However, only for the input image correctly classified as 'FNA-not-required', do the heatmaps display highly activated areas where the nodule is located. For the rest of the cases, across the different algorithms, the decision of the model is based on the incorrect part of the input images. Moreover, in all cases, the ScoreCAM completely fails to provide any information.

5.6.4. Heatmaps using attention weights

Interpretability is less straightforward for the heatmaps directly generated from the weights of the spatial attention module. As demonstrated in Figure 12, areas containing meaningful spatial information are dispersed throughout the entire image, intermingled with areas of less significant information. Nevertheless, it is observable that, for the case where the input image was correctly classified as 'FNA-not-required', this activated area is still situated around the position of the nodule.

6. Discussion

6.1. Optimization and refinement of the network

The initial phase of the study focused on optimizing the baseline model, which was critical in establishing a benchmark for subsequent experiments. This process was a cornerstone in the development of a more refined and efficient network, as it involved varying the hyperparameters and choosing the configuration that yielded the lowest validation loss.

Building on this, the results accentuated the substantial advantage of employing pre-trained weights. They provided a robust foundation for the model, improved performance metrics, and mitigated the risk of overfitting. This insight reemphasizes the significance of pre-training in the network setup.

Following the experiments performed, the addition of extra hidden layers showed a positive influence on network performance. The assumption that more layers could potentially degrade the model's performance was negated. An optimal configuration with four hidden layers was identified that exhibited lower validation loss, particularly on the "Baseline" dataset. However, it also raised questions about the potential limitations of added complexity in the classifier when applied to the "Enlarged" dataset.

6.2. Contrasting performance of Mono-Input and Multi-Input models

The study explored mono-input and multi-input models, providing valuable insights, especially considering the dataset characteristics and employed

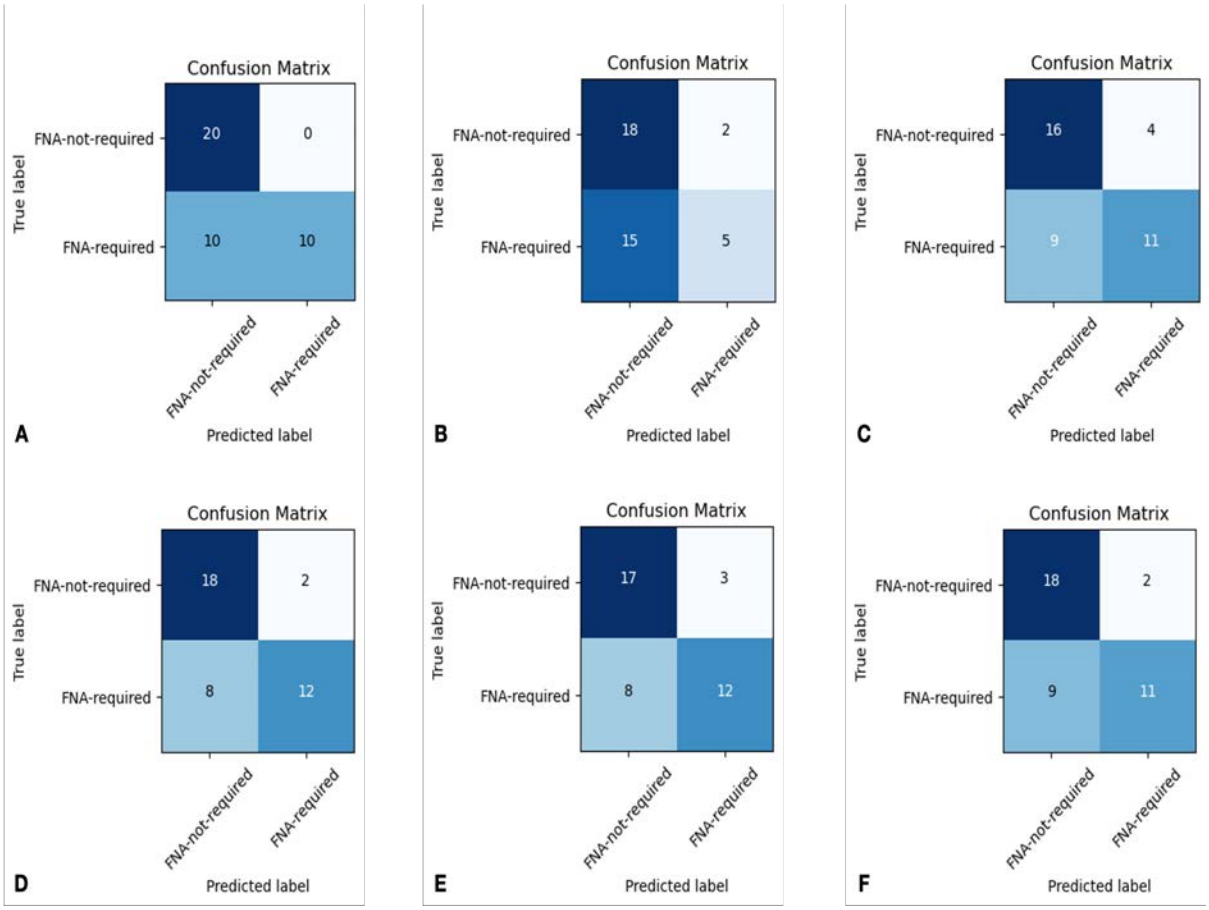


Figure 7: Confusion matrices representing the binary classification results of thyroid nodules on the test set for the mono-input network. **A**: The model that achieved the lowest validation loss across the five folds during cross-validation on the Baseline dataset. **B**: The model trained on the entire Baseline dataset. **C**: An ensemble model that uses the weights of each of the models with the lowest validation loss from the five folds during cross-validation on the Baseline dataset. **D**: The model that achieved the lowest validation loss across the five folds during cross-validation on the Enlarged dataset. **E**: The model trained on the entire Enlarged dataset. **F**: An ensemble model that uses the weights of each of the models with the lowest validation loss from the five folds during cross-validation on the Enlarged dataset.

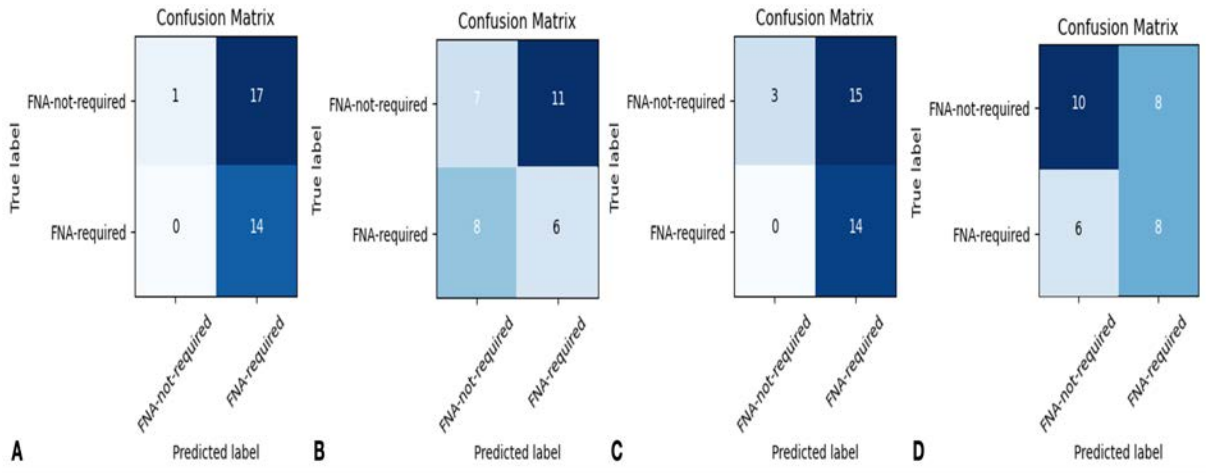


Figure 8: Confusion matrices of the binary classification of thyroid nodules on the test set for the multi-input network. **A** represents the multi-input model utilizing input-level fusion, **B** represents the multi-input model employing feature-level fusion, **C** represents the multi-input model utilizing intermediate-level fusion, and **D** represents the multi-input model employing decision-level fusion.

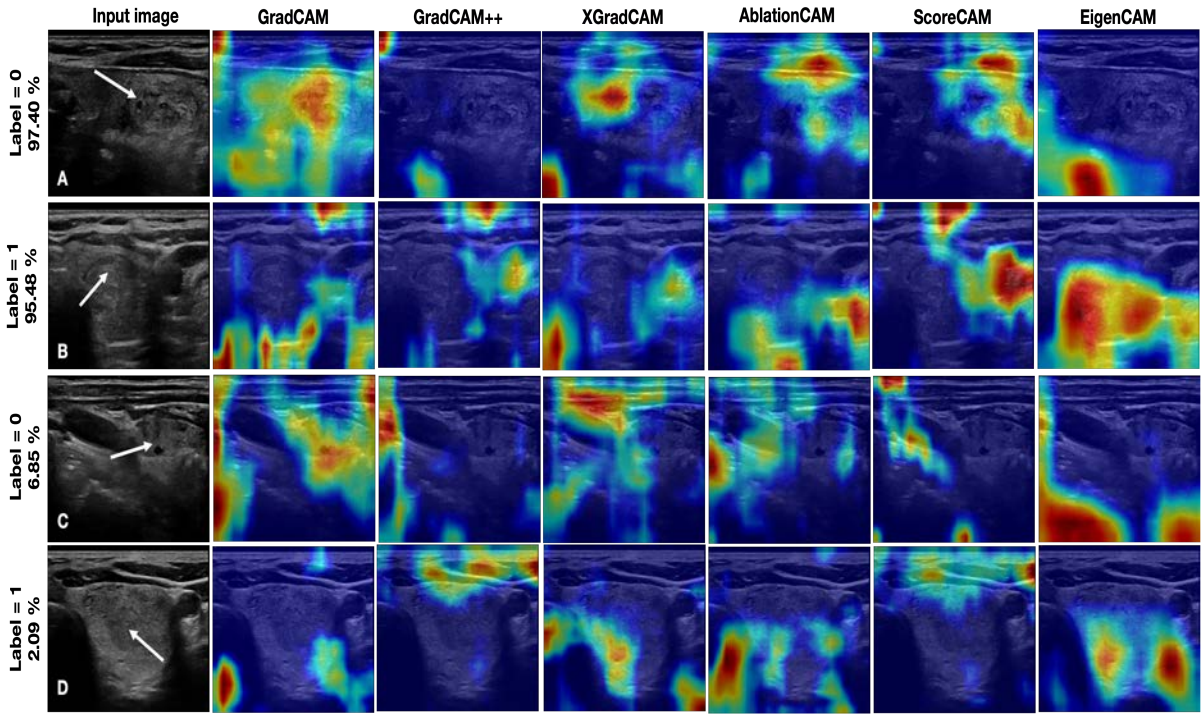


Figure 9: Comparison of heatmaps generated using different CAM algorithms, extracting the information from the final layer of the backbone. The first element in each row represents the input images, to the left of which the ground-truth label and the model's class probability are displayed. **A**: Image correctly classified as 'FNA-not-required'. **B**: Image correctly classified as 'FNA-required'. **C**: Image incorrectly classified as 'FNA-required'. **D**: Image incorrectly classified as 'FNA-not-required'. The white arrows in the first column indicate the nodule's location in each case.

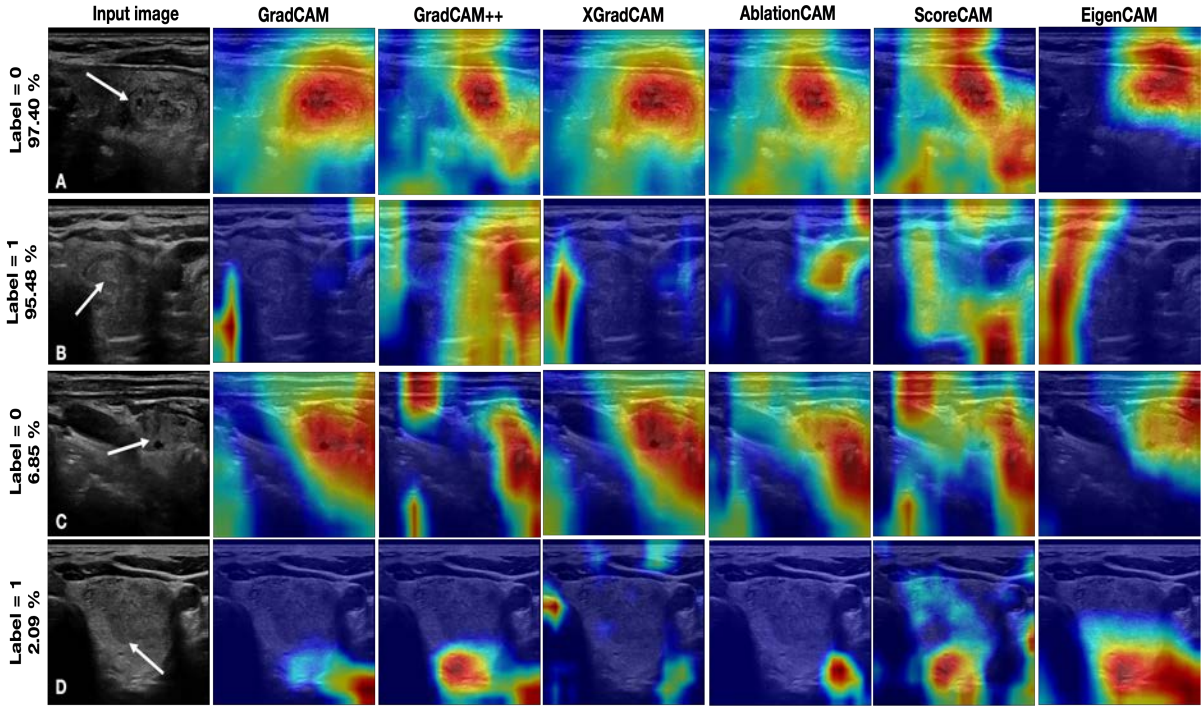


Figure 10: Comparison of heatmaps generated using different CAM algorithms, extracting the information from the spatial attention module layer of the overall architecture. The first element in each row represents the input images, to the left of which the ground-truth label and the model's class probability are displayed. **A**: Image correctly classified as 'FNA-not-required'. **B**: Image correctly classified as 'FNA-required'. **C**: Image incorrectly classified as 'FNA-required'. **D**: Image incorrectly classified as 'FNA-not-required'. The white arrows in the first column indicate the nodule's location in each case.

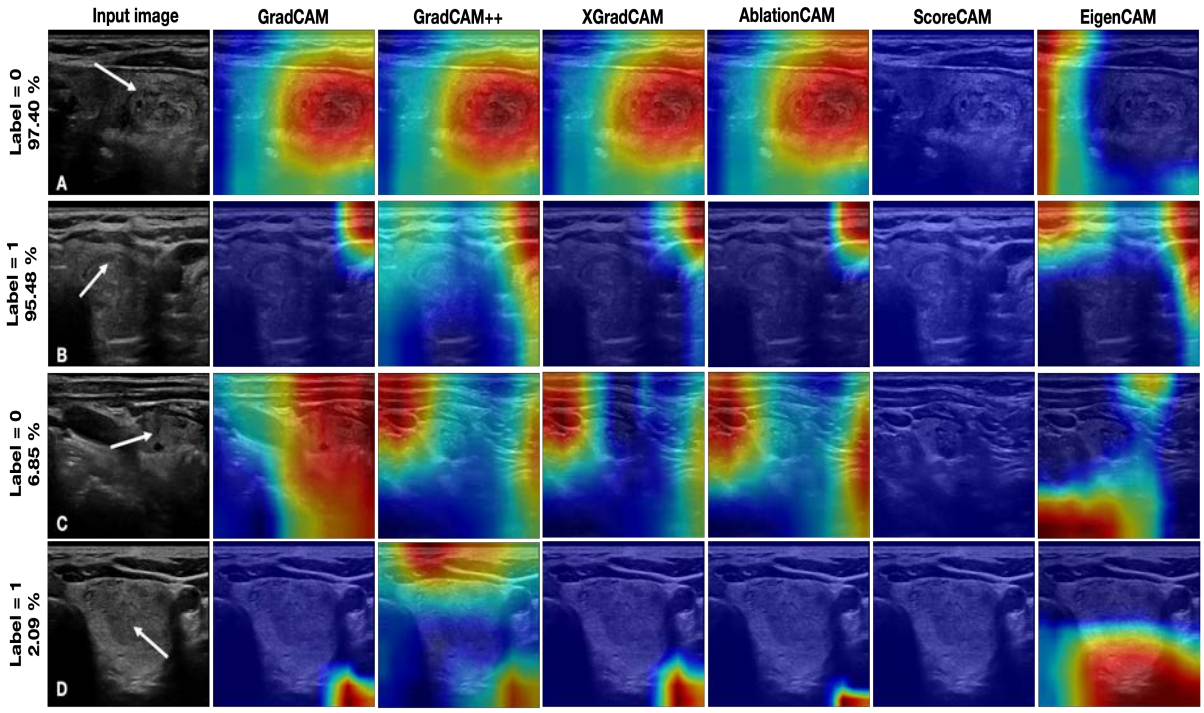


Figure 11: Comparison of heatmaps generated using different CAM algorithms, extracting the information from the final convolutional layer of the overall architecture. The first element in each row represents the input images, to the left of which the ground-truth label and the model's class probability are displayed. **A**: Image correctly classified as 'FNA-not-required'. **B**: Image correctly classified as 'FNA-required'. **C**: Image incorrectly classified as 'FNA-required'. **D**: Image incorrectly classified as 'FNA-not-required'. The white arrows in the first column indicate the nodule's location in each case.

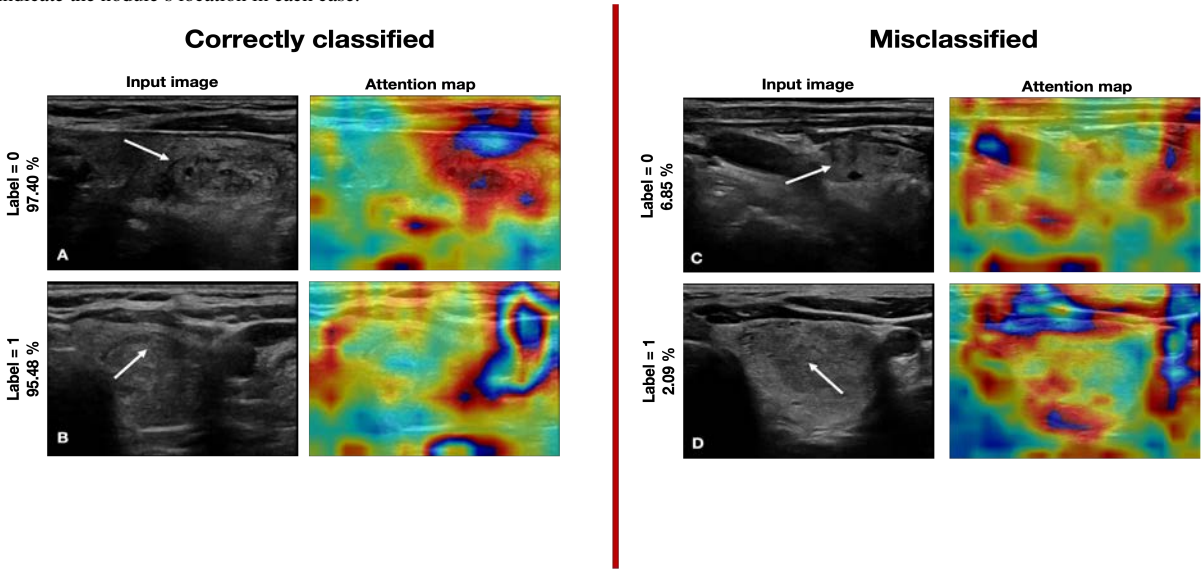


Figure 12: Heatmaps generated using the weights extracted from the spatial attention module. On the left are images correctly classified by the model, and on the right are images incorrectly classified by the model. The probabilities assigned by the model for each sample to belong to the corresponding class are also displayed. The white arrows in the first column indicate the nodule's location in each case.

methodology. The problem complexity was amplified in this setting, affirming the significance of the dataset construction and data quality in DL model success.

Interestingly, the mono-input model displayed superior performance over the multi-input model, even under different fusion strategies. Typically, multiple inputs boost the neural networks' learning capacity, allowing them to capture complex patterns in data. However, in the context of this study, additional information seemed to increase the noise rather than im-

prove class distinction. Furthermore, the inclusion of multiple inputs reduced the number of cases by two and increased the complexity of the data.

Upon deeper analysis of the performance characteristics of the mono-input and multi-input models, a striking difference was observed. High specificity was exhibited by the mono-input models, indicating proficiency in correctly identifying samples that were indeed negative, or 'FNA-not-required'. Conversely, high sensitivity was demonstrated by the multi-input models, signifying their capability in accurately recognizing positive samples, those labeled as 'FNA-required'.

These findings highlight the trade-off between specificity and sensitivity in the mono-input and multi-input models. Achieving a balance between the two metrics is crucial to ensure accurate classification and reduce the risk of false interpretations in medical scenarios.

6.3. Evaluating the efficacy of fusion strategies

When comparing different fusion strategies, decision-level fusion demonstrated the best performance. By performing fusion at the decision level, the network can rely on individual decisions made based on different data subsets. This strategy likely contributed to a more comprehensive understanding of the application, which could lead to more accurate overall decisions.

6.4. Impact of dataset characteristics and model performance metrics

The impact of dataset characteristics on model performance was evident in the results of the binary classification task. Models trained on a subset of the data (Models A and D) outperformed those trained on the entire dataset (Models B and E), as shown in Table 7. This performance difference highlights the importance of balanced class distributions in the training data.

Interestingly, the optimal performance was achieved by models A and D, which were trained on subsets with more balanced class distributions. This suggests that when the model was exposed to a dataset with a more balanced representation of classes during training, it was better able to generalize and perform well during inference. These findings emphasize the critical role that class distribution plays in training robust and high-performing models.

It is worth noting that the "Baseline" dataset, which served as the initial reference dataset in this study, was expanded to create the "Enlarged" dataset. The expansion involved combining the original "Baseline" dataset with newly acquired images, exclusively sourced from the Hospital of Bastia. It is important to highlight that the images from the Hospital of Bastia were obtained using a different scanner compared to the Hospital of Dijon. The introduction of these new images, along with potential variability in image quality and characteristics, may have contributed to the observed decline in performance when training on the "Enlarged" dataset

Overall, these results underscore the importance of careful dataset construction, including considerations of class distribution and potential variations in image sources, in order to train robust models with optimal performance.

6.5. Explainability through heatmaps

In the analysis of explainability through heatmaps, it was observed that the interpretability of the models varied based on the layer and method used. The last layer of the backbone did not provide meaningful information about the nodule's localization, as shown in Figure 9. However, GradCAM successfully indicated the correct nodule localization for 'FNA-not-required' predictions, while EigenCAM, despite displaying a larger activation area, failed to accurately pinpoint the nodule's location for 'FNA-required' cases.

Additionally, the spatial attention module within the CBAM played a crucial role in correctly identifying the nodule's position for 'FNA-not-required' cases, as demonstrated in Figure 10. This finding was further reinforced by the effective highlighting of the nodule's position through GradCAM and XGradCAM. In contrast, ScoreCAM and EigenCAM extended their activations to areas outside the nodule region, and none of the algorithms successfully localized the nodule for 'FNA-required' classifications. Interestingly, even in cases of misclassification, the activated areas still exhibited a close alignment with the actual position of the nodule.

In cases where the model made accurate predictions but the heatmaps did not accurately highlight the nodule's location, it is important to consider that the model may be utilizing other features or contextual information surrounding the nodule for classification. These features may not be visually apparent in the generated heatmaps. The model could be capturing subtle patterns or contextual cues that are learned during training. Therefore, although the heatmaps may not show specific activation in the nodule region, the model could still be leveraging other relevant features to make the correct decision. This emphasizes the DL models' ability to capture and utilize complex information in decision-making, even when it is not explicitly evident in the generated heatmaps.

On the other hand, when evaluating the performance of different CAM algorithms across different cases and target layers, GradCAM consistently demonstrated the best performance. The heatmaps generated for each input image remained consistent across different stages in the network.

Lastly, the interpretability of the models was less clear when considering the heatmaps generated directly from the spatial attention module's weights,

as shown in Figure 12. Although significant spatial information appeared scattered throughout the entire image, the nodule's location could still be discerned in correctly classified 'FNA-not-required' cases.

6.6. Comparison with the previous internship

In comparison to the previous internship, it is important to highlight the improvement in the performance of the Attention-Densenet121 network in this study. The confidence of the network in its predictions was enhanced by optimizing the loss metric of the model. It is worth noting that the previous internship exclusively focused on exploring the mono-input configuration, making a direct performance comparison between the mono-input and multi-input networks unfeasible. However, within the context of the mono-input setting, it can be observed that the network in this study achieved a high level of specificity, which was the primary objective. Nonetheless, this improvement in specificity came at the expense of a slight decrease in sensitivity.

7. Future work

To further enhance the performance of the Mono-input and Multi-input Attention-Densenet121 networks presented in this study, it is crucial to determine the optimal configuration of the dataset for training both network settings. Finding the most effective combination of training data sources, image characteristics, and DA techniques can potentially improve the models' classification accuracy and generalization capability.

Additionally, identifying the XAI algorithm that provides more precise information about nodule localization is of utmost importance. Developing an algorithm that can serve as a second opinion in a medical setting relies on accurate and interpretable nodule localization. This would provide valuable support to medical professionals in making informed decisions.

Moreover, given the limitations of the TIRADS system in determining relevant image features, there is a critical need for interpretable representations that can capture and explain the key characteristics of thyroid nodules. Techniques such as unsupervised learning or disentangled representation learning hold promise in extracting meaningful and interpretable features from ultrasound images. By gaining a deeper understanding of the underlying factors associated with different types of thyroid nodules, more accurate and explainable diagnostic models can be developed. These models can provide valuable insights to medical professionals, ultimately improving decision-making and enhancing patient care in the field of thyroid nodule classification.

8. Conclusions

In conclusion, the study involved optimizing and refining the network, which resulted in improved performance. Mono-input models demonstrated superior performance compared to multi-input models. Among the fusion strategies, decision-level fusion showed the best results. Models trained on a balanced subset of the data outperformed those trained on the entire dataset. The interpretability of the models varied depending on the method and layer used. These findings contribute valuable insights into the field of DL models for thyroid nodule classification, highlighting the importance of model optimization, dataset characteristics, and interpretability techniques for future research and advancements in this domain.

Acknowledgments

First and foremost, I would like to express my profound gratitude to Dr. Catherine Vuillemin and Dr. Serge Angiolini from the Hospital of Bastia, France, as well as Dr. Perrine Buffier, Dr. Elodie Crevisy, Dr. Amandine Nguyen, Dr. Marie-Paule Monnier Meteau, and Dr. Pauline Legris, from the Hospital of Dijon, for their invaluable contribution in collecting the images for the dataset used in this study. The availability of high-quality data is paramount for the development of high-performance DL models, and this study would not have been possible without their assistance.

I am also deeply thankful to my supervisors, Dr. Sarah Leclerc and Dr. Alain Lalande, for their unwavering trust in me, invaluable time, and guidance throughout this demanding project.

Furthermore, I would like to express my heartfelt gratitude to my family, especially my parents, for being the bedrock of my education and providing unwavering support throughout my academic journey.

Lastly, I extend my sincere appreciation to the European Union and the Erasmus Mundus Joint Master Degree in Medical Imaging and Applications (MAIA) for granting me the opportunity to fulfill one of my lifelong dreams

References

Biskup, D., 2008. A state-of-the-art review on scheduling with learning effects. European Journal of Operational Research 188, 315–329.

Buda, M., Wildman-Tobriner, B., Hoang, J.K., Thayer, D., Tessler, F.N., Middleton, W.D., Mazurowski, M.A., 2019. Management of thyroid nodules seen on us images: deep learning may match performance of radiologists. Radiology 292, 695–701.

Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N., 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, in: 2018 IEEE winter conference on applications of computer vision (WACV), IEEE. pp. 839–847.

Chi, J., Walia, E., Babyn, P., Wang, J., Groot, G., Eramian, M., 2017. Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. Journal of digital imaging 30, 477–486.

Cibas, E.S., Ali, S.Z., 2009. The bethesda system for reporting thyroid cytopathology. Thyroid 19, 1159–1165.

Cooper, D.S., 2009. American thyroid association (ata) guidelines taskforce on thyroid nodules and differentiated thyroid cancer. revised american thyroid association management guidelines for patients with thyroid nodules and differentiated thyroid cancer. Thyroid 19, 1167–1214.

Dasmehdi, Xtr, 2017. Ddti thyroid ultrasound images. URL: https://www.kaggle.com/datasets/dasmehdixtr/ddti-thyroid-ultrasound-images. last access: 2023-03-04.

Deng, P., Han, X., Wei, X., Chang, L., 2022. Automatic classification of thyroid nodules in ultrasound images using a multi-task attention network guided by clinical knowledge. Computers in Biology and Medicine 150, 106172.

Fu, R., Hu, Q., Dong, X., Guo, Y., Gao, Y., Li, B., 2020. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. arXiv preprint arXiv:2008.02312 .

Gildenblat, J., contributors, 2021. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam. Last access: 2023-05-27.

Grant, E.G., Tessler, F.N., Hoang, J.K., Langer, J.E., Beland, M.D., Berland, L.L., Cronan, J.J., Desser, T.S., Frates, M.C., Hamper, U.M., et al., 2015. Thyroid ultrasound reporting lexicon: white paper of the acr thyroid imaging, reporting and data system (tirads) committee. Journal of the American college of radiology 12, 1272–1279.

Guo, M.H., Xu, T.X., Liu, J.J., Liu, Z.N., Jiang, P.T., Mu, T.J., Zhang, S.H., Martin, R.R., Cheng, M.M., Hu, S.M., 2022. Attention mechanisms in computer vision: A survey. Computational Visual Media 8, 331–368.

Hayat, M.J., Howlader, N., Reichman, M.E., Edwards, B.K., 2007. Cancer statistics, trends, and multiple primary cancer analyses from the surveillance, epidemiology, and end results (seer) program. The oncologist 12, 20–37.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE international conference on computer vision, pp. 1026–1034.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Hoang, J.K., Middleton, W.D., Farjat, A.E., Teefey, S.A., Abinanti, N., Boschini, F.J., Bronner, A.J., Dahiya, N., Hertzberg, B.S., Newman, J.R., et al., 2018. Interobserver variability of sonographic features used in the american college of radiology thyroid imaging reporting and data system. American Journal of Roentgenology 211, 162–167.

Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.R., Samek, W., 2022a. xxai-beyond explainable artificial intelligence, in: xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers, Springer. pp. 3–10.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H., 2019. Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9, e1312.

Holzinger, A., Saranti, A., Molnar, C., Biecek, P., Samek, W., 2022b. Explainable ai methods-a brief overview, in: xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers, Springer. pp. 13–38.

Kong, M., Guo, Q., Zhou, S., Li, M., Kuang, K., Huang, Z., Wu, F., Chen, X., Zhu, Q., 2022. Attribute-aware interpretation learning for thyroid ultrasound diagnosis. Artificial Intelligence in Medicine 131, 102344.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J., 2019. On the variance of the adaptive learning rate and beyond. arXiv preprint arXiv:1908.03265 .

Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 .

Manh, V.T., Zhou, J., Jia, X., Lin, Z., Xu, W., Mei, Z., Dong, Y., Yang, X., Huang, R., Ni, D., 2022. Multi-attribute attention network for interpretable diagnosis of thyroid nodules in ultrasound images. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control 69, 2611–2620.

Meyes, R., de Puiseau, C.W., Posada-Moreno, A., Meisen, T., 2020. Under the hood of neural networks: Characterizing learned representations by functional neuron populations and network ablations. arXiv preprint arXiv:2004.01254 .

Muhammad, M.B., Yeasin, M., 2020. Eigen-cam: Class activation map using principal components, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE. pp. 1–7.

Ramaswamy, H.G., et al., 2020. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 983–991.

Schnadig, V.J., 2018. Overdiagnosis of thyroid cancer: is this not an ethical issue for pathologists as well as radiologists and clinicians? Archives of pathology & laboratory medicine 142, 1018–1020.

Seeland, M., Mäder, P., 2021. Multi-view classification with convolutional neural networks. Plos one 16, e0245230.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, pp. 618–626.

Singh, A., Sengupta, S., Lakshminarayanan, V., 2020. Explainable deep learning models in medical image analysis. Journal of Imaging 6, 52.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9.

Szeghalmy, S., Fazekas, A., 2023. A comparative study of the use of stratified cross-validation and distribution-balanced stratified cross-validation in imbalanced learning. Sensors 23, 2333.

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X., 2020. Score-cam: Score-weighted visual explanations for convolutional neural networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp. 24–25.

Woo, S., Park, J., Lee, J.Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV), pp. 3–19.

Yang, J., Shi, X., Wang, B., Qiu, W., Tian, G., Wang, X., Wang, P., Yang, J., 2022. Ultrasound image classification of thyroid nodules based on deep learning. Frontiers in Oncology , 3545.

Zhang, D., Nayak, R., Bashar, M.A., 2021. Exploring fusion strategies in deep learning models for multi-modal classification, in: Data Mining: 19th Australasian Conference on Data Mining, AusDM 2021, Brisbane, QLD, Australia, December 14-15, 2021, Proceedings, Springer. pp. 102–117.

Zhang, J., He, T., Sra, S., Jadbabaie, A., 2019. Why gradient clipping accelerates training: A theoretical justification for adaptivity. arXiv preprint arXiv:1905.11881 .

Zhu, Y., Fu, Z., Fei, J., 2017. An image augmentation method using convolutional network for thyroid nodule classification by transfer learning, in: 2017 3rd IEEE International Conference on Computer and Communications (ICCC), IEEE. pp. 1819–1823.

Medical Imaging and Applications

Master Thesis, June 2023

Supervised Automatic Segmentation of Foot Bones in 3D CT scans

Leticia Itzel Rivera Contreras,
Supervised by: Sergi Pujades Rocamora[1], Julien Pansiot[1]

INRIA, Grenoble, France[1]

leticia-itzel.rivera-contreras@inria.fr, sergi.pujades-rocamora@inria.fr, julien.pansiot@inria.fr

Abstract

Rheumatoid arthritis is a condition that affects thousands of people worldwide. The treatments that exist today are not a complete cure but only an aid to slow the progression of arthritis as well as relieve some symptoms. According to the patients, pain is one of the most prevalent symptoms, which leads to emotional and quality of life consequences. On the other hand clinical studies have demonstrated that insoles could be useful for the improvement of pain. Nevertheless, experiments in this area are still unexplored, so in order to carry out the development of the insoles, it is necessary to resort to computational models that provide the finite element analysis with accurate and precise data on the identification and location of the foot bones for the design of the prototype. The first step for a computational model is to use modern tools based on deep learning to obtain a proper segmentation of the foot bones. In this work a method based on a 3D U-Net with Attention that contributes to the automatic segmentation of the foot bones is proposed. In the first experiments, a decent segmentation of the posterior foot bones has been obtained, so further tests were done to obtain a more defined segmentation on the forefoot bones. The trained models have been tested on two different datasets, one coming from INRIA and the other from CHU Saint Etienne. The last one completely unknown data for the model. The evaluation demonstrated that the main general goal was achieved in a qualitative and quantitative way.

Keywords: Automated segmentation, Supervised segmentation, Foot bones, 3D Attention UNET

1. Introduction

The foot is a remarkable structure composed of numerous bones that work together to support body weight, provide stability, and facilitate movement (Parvizi, 2021). Inside the foot are around 30 joints, and more than 100 ligaments, tendons and muscles (Hardy and Snaith, 2011). Accurate knowledge of foot bone morphology and spatial relationships is crucial for diagnosing and treating various foot disorders, injuries, and deformities (Coughlin and Mann, 2014).

The foot consists of three major regions: the hind, mid, and forefoot. The hindfoot includes the talus and calcaneus bones, forming the ankle joint and providing a stable base for the foot (Coughlin and Mann, 2014). The midfoot comprises the navicular, cuboid, and cuneiform bones, giving the arch support and flexibility. The forefoot comprises the metatarsals and phalanges, forming the toes and enabling weight-bearing, propulsion, and balance during locomotion (Standring, 2016). In the figure [Fig.1], it can be observed the three parts of the foot mentioned above.

Each bone in the foot exhibits distinct anatomical features and variations, including size, shape, articulations, and bony prominences. For example, the talus bone is a tarsal bone that articulates with the tibia and fibula, forming the ankle joint. It plays a crucial role in transmitting forces between the leg and foot (Standring, 2016). The calcaneus bone, also known as the heel bone, supports body weight and is an attachment site for muscles and ligaments (Coughlin and Mann, 2014).

Many degenerative conditions affect bones. One of them is arthritis which cannot be considered a single disease. The term refers to joint discomfort or illness, and there are different varieties of arthritis and disorders associated with it. However, the current medicine does

Figure 1: Three main regions of the foot. Midfoot in green, hindfoot in pink and forefoot in blue

not have a cure for this pathology, thus, the treatments are based on symptom relief (Pelt, 2012).

Rheumatoid arthritis (RA) is a chronic autoimmune disease characterized by joint inflammation, pain, and progressive joint damage (Aletaha and Smolen, 2018). It affects millions worldwide, including a significant population in France with a prevalence of 0.5% (Egan et al., 2001). Hands and feet are commonly affected by RA, a type of peripheral polyarthritis that compromises foot function and is accompanied by changes in plantar pressure and gait abnormalities. Patients are negatively affected, which results in foot pain, instability, difficulty walking, and a lower quality of life (Turner et al., 2008).

RA treatment has improved due to the advent of anti-rheumatic drugs since patients with severe inflammatory and destructive illnesses have been reduced. However, pain relief remains a significant concern for many patients as the emotional and quality of life consequences are significantly affected by this symptom (Vergne-Salle et al., 2020). Some studies have shown that footwear is essential in treating RA, as tight-fitting shoes can lead to high pressures and consequent pain. For this reason, foot orthoses have been prescribed to improve lifestyle quality by reducing the load in regions where pain may be concentrated (Kelly et al., 2021). Nevertheless, their effectiveness is still unclear.

Furthermore, biomechanical researchers investigate foot bones' mechanical properties and function to develop improved footwear designs and understand gait abnormalities. By using therapeutic footwear and personalized foot orthoses, foot pain and impairment can be decreased (Guillemin et al., 2005). Orthotic insoles are frequently employed to treat RA patients and increase their functional abilities. Although wearing insoles is commonly associated with pain alleviation, systematic evaluation of the mechanisms involved in this treatment is lacking. Due to the difficulty of conducting such studies in a clinical setting, insole design and its relationship to internal effects such as joint pressure and soft tissue deformations have yet to be examined (Egan et al., 2001).

In other words, understanding foot pain still has several limitations due to the need for methods to study the sources of pain and the complex structure of the foot that does not allow us to understand specific mechanisms involved in pain. The finite Element Method (FEM) offers a powerful computational approach to tackling this issue. It can model the mechanical response of the foot to different stimuli and investigate the distribution of stresses within different tissues allowing a deeper understanding of structure abnormalities, tissue degeneration, and other factors involved in pain. FEM also facilitates the identification of specific regions that experience excessive stress or strain, helping localize potential pain sources (Chen et al., 2015).

On the other hand, computational modeling provides a further understanding of how loading affects the soft tissue and joints internally. There are different tools to model the foot, in this case, the bones. One widely used technique in orthopedics for surgery planning, biomechanical analysis, and the development of custom-fit prosthetics is segmentation (Pham et al., 2000). In conjunction with FEM, segmentation plays a crucial role in precisely localizing pain sources within the foot because it provides accurate and detailed information about the geometry and boundaries of the analyzed structures.

In the same way, segmentation has shown promising results in various medical modalities. Computed tomography (CT), magnetic resonance imaging (MRI), and X-rays can provide valuable insights into the foot bones' structural characteristics and spatial relationships. Using these imaging techniques, one can precisely identify, segment, and measure each bone in a 2D or 3D image. Additionally, imaging facilitates the detection of pathological conditions, deformities and fractures. Podiatrists utilize foot bone analysis to diagnose and manage conditions like arthritis, flatfoot, or bunion deformities (Zhang et al., 2014).

The segmentation process involves identifying and separating individual bones within the foot. Supervised segmentation leverages labeled training data to learn a segmentation model that can accurately delineate bones in CT scans. The model can then be applied to unseen scans to automate segmentation. However, due to the complex anatomical structures and variations in bone shapes, manual segmentation of foot bones is a time-consuming and challenging task for radiologists and clinicians (Wang and Hazlehurst, 2018). For this rea-

son, automatic segmentation has become the solution to this problem. Machine learning methods and convolutional neural networks (CNN) are required for this type of segmentation.

The primary objective of this paper is to develop and evaluate a supervised segmentation framework for accurately segmenting foot bones in 3D CT scans, focusing on patients with rheumatoid arthritis in France. It is worth mentioning that this task corresponds to the first stage of a project called INORA. The main goal of INORA is to investigate the mechanisms of action of shoes and orthotic insoles using patient-specific computational biomechanical models to suggest a reasonable approach to their design. On a more fundamental view, these models will aid in discovering mechanical determinants of pain reduction, allowing for patients' long-term well-being.

Hence, the contributions of this work are as follows:

- Creation of a dataset with their respective Ground Truth (GT) of healthy patients.
- Generation of data cubes, which are used as input to the network.
- Application of a 3D U-NET with attention blocks to perform binary supervised automated segmentation.
- A Weighted loss function to discriminate background pixels.
- Training focused on the Field of View (FOV)
- Obtention of 3D meshes with different segmentation methods: overlapping, majority voting, and sliding.
- 3D meshes with all and individual structures of the foot
- Evaluation of the predictions with statistic methods to know the model's performance on each image.

2. State of the art

Medical image segmentation plays a crucial role in various clinical applications, enabling accurate diagnosis, treatment planning, and disease monitoring. Over the years, researchers have made significant advancements in the field, proposing innovative techniques to address the challenges associated with segmenting medical images. This section provides an overview of some techniques in medical image segmentation. It will focus on deep learning methodologies and algorithms based on supervised segmentation of the spine and vertebrae due to the limited information available on foot segmentation.

For biomedical image segmentation tasks, the network known for excellence and its superior performance results is the U-NET architecture presented by (Ronneberger et al., 2015). U-NET is a network based on

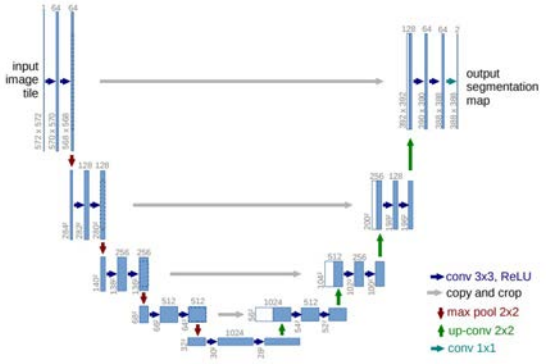


Figure 2: UNET architecture for medical segmentation. Courtesy of (Ronneberger et al., 2015)

CNN networks. It is built on an encoder-decoder structure where the encoder learns hierarchical representations of the input image through convolutional layers, capturing low-level and high-level features. The decoder then utilizes upsampling and skip connections to recover spatial information and generate a pixel-wise segmentation map. U-NET's architecture shown in figure [Fig.2], with its contracting and expanding pathways, has proven effective in capturing context and local details, enabling accurate and efficient segmentation.

On the other hand, in his paper based on the challenge VerSe 2019, (Payer et al., 2020) proposes a three-step method to perform vertebrae localization and segmentation. It consists of using two CNNs, one with a coarse input resolution, to predict the approximate location of the spine. And the other at a higher resolution for the multiple landmark localization and identification of individual centroids. Finally, the segmentation CNN performs a binary image segmentation of each localized vertebra in the highest resolution.

Furthermore,they use a single network for segmenting the vertebra because each one is individually separated from the background. Since each vertebra has an independent segmentation, the network must know which vertebrae are in the input volume to segment. Thus, they crop the region surrounding the localized vertebra from the CT scan to the center of the reduced image. During inference, they utilize the ground-truth vertebra location throughout training and the predicted vertebra. Besides, they generate an image of a Gaussian heatmap centered on the vertebra location. Both the cropped and heatmap images are sent into the segmentation U-Net.

They adapt the U-NET of (Ronneberger et al., 2015) for the binary segmentation to execute average instead of max pooling and linear up-sampling instead of transposed convolutions. The network has five levels with a kernel size of [3x3x3] and 64 filters and is set up to predict a single output volume. The vertebrae segmentation method is shown in [Fig.3]. The results obtained on the test sets for the segmentation have good performance

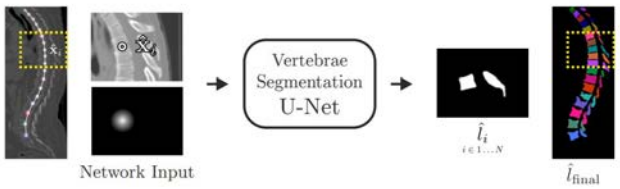


Figure 3: General overview of vertebrae segmentation using U-NET method proposed by (Payer et al., 2020)

compared to others, reaching a method with good generalization capabilities and winning the VerSe 2019 challenge.

(Meng et al., 2023) proposed a novel framework that combines graph optimization and an anatomical consistency cycle to improve the accuracy and robustness of vertebrae analysis. These techniques are applied between the localization, segmentation, and identification tasks. Graph optimization is used to refine localization and segmentation results by leveraging spatial relationships between neighboring vertebrae. At the same time, incorporating an anatomical consistency cycle ensures anatomical accuracy throughout the analysis process. For the vertebrae segmentation, they computed a binary segmentation of the entire spine that helps locate the region of interest. This step allows the estimation of the locations of individual vertebrae. With this information, the individual vertebrae segmentation masks are calculated. Similarly to (Payer et al., 2020), they use three steps strategy of localization, identification, and segmentation with the difference that they introduce an anatomic consistency cycle. This cycle ensures that the results are refined iteratively.

For the segmentation task, they use a 3D Attention U-NET. The base architecture used is the 3D U-NET of (Ronneberger et al., 2015), but with the segmentation module in (Payer et al., 2020) and adding attention blocks (Oktay et al., 2018) embedded in the skip connections. The network's input is a 3D CT image patch, and its output is a probability mask for each voxel, indicating whether or not it belongs to the bone. The overview of the method is visualized in the figure [Fig.4].

Furthermore, the vertebrae are identified using locations and segmentation masks. A VGG network is used for this task considering both local and global contexts. The classification is made in two stages: first, predicting the spine level (cervical, thoracic, or lumbar) and then predicting the identity within the group.

The results obtained achieved state-of-the-art execution on the VerSe challenge. Furthermore, the method generalizes well in unseen data.

Inspired by the work of (Meng et al., 2023), a new approach to segmenting the foot bones is developed. The method is intended to help the finite element analysis in order to design insoles that allow dealing with the pain provoked by Rheumatoid Arthritis.

The proposed framework consists firstly in achieving an automatic binary segmentation of CT volumes in two different perspectives of the foot, which does not consider prior knowledge. It should also be mentioned that the phalanges were not taken into account. Moreover, it incorporates a loss function with weights to discard pixels with no bone, an approach centered on the FOV of the scan, and four models trained with the four different groups of bones to analyze the complex structures found in segmentation results. Additionally, a prediction aggregation is included in three modes: Overlapping, sliding, and central voting to generate the final masks.

3. Material and methods

3.1. Dataset

The dataset used in this paper is building-labeled. The data was provided by Centre Inria de l'Université Grenoble Alpes. It includes 3D CT scans of 11 healthy patients in-vivo and their respective meshes segmented by an image processing method. Each patient has two images, one corresponding to the anterior zone of the foot and the other to the posterior location of the foot. In every scan, different bones can be appreciated. Then, they merged to form the entire foot. Some images from this dataset can be observed in [Fig.5]

3.1.1. Ground Truth creation

For supervised learning purposes, an annotated dataset was created. The 22 volumes were already segmented into 14-15 bones and saved in an STL format. For the reconstruction and voxelization of these STL files, 3D slicer was used. The first step of this process was to load the 14 or 15 segments separately to join them into a single segmentation later and create a binary label map with a single layer. The map is then selected, and the volume of the original image is used as a reference to ensure that all CT characteristics are considered to create a new one. Finally, it is exported as a NIFTI image. This procedure was repeated 22 times, one for each image of each patient. At that point, the dataset had been labeled and was ready for usage.

3.2. Pre-processing

Once the dataset was created, it was observed and cleaned. The image was adapted to the field of view

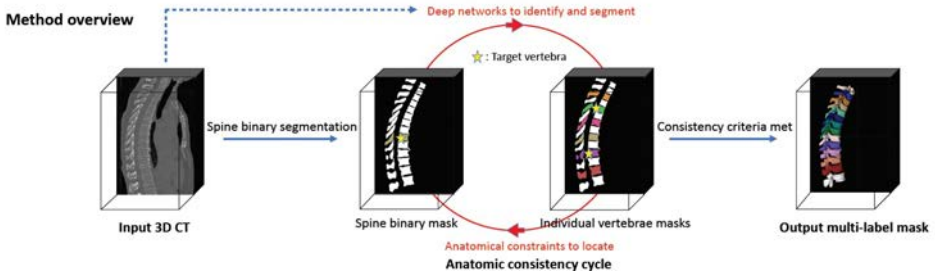


Figure 4: Vertebrae segmentation: overview method using 3D U-Net with Attention. Courtesy of (Meng et al., 2023)

(FOV), resulting in the removal of the CT table after the obtention of the first results. However, this adjustment helped to focus more precisely on the area to be segmented.

3.2.1. Test dataset

For testing purposes, another dataset was used. This dataset comprises 3D CT scans of three healthy patients ex-vivo from CHU Saint-Etienne and does not contain labels. These scans are in a DICOM format, so a conversion to NIFTI format was performed for better data manipulation. The image was resized to fit the field of view in the training dataset.

Figure 5: The first four images belong to the anterior foot, and the last two to the hindfoot. The cuneiform bones, the navicular, and the cuboid can be appreciated on the left. In the middle, the five metatarsus and the tibia, fibula, calcaneus, and talus are shown on the right.

3.3. Data preparation

3.3.1. Cube extraction

A cube extraction strategy was employed because the network only accepts fixed-size input, and the CT scan input has an arbitrary size. All the dataset was preprocessed. The cube extraction was done taking into account various cube sizes and strides. The overlap was created by making the stride smaller than the cube size. Furthermore, volume padding was considered, with a value of -1000, the value of the air in Hounsfield units for the images, and a value of 0 for the GT. The scans were also down-sampled due to decreased voxel spacing, although trials were performed with the original voxel spacing. The following combinations of three cube sizes and strides with two different voxel spacing were tried to obtain the optimal network settings:

1. 96x96x96 and 48 with 0.37mm and 1.0mm
2. 64x64x64 and 32 0.37mm and 1.0mm
3. 32x32x32 and 16 0.37mm and 1.0mm

Once the cubes were obtained, a viewer was created to visualize them. See the images from different views to check if the extraction was successful was possible with this tool. An example is shown in the [Fig.6].

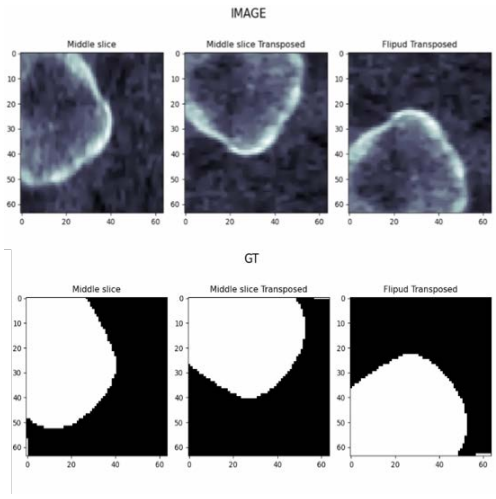


Figure 6: Example of a cube with three different view perspectives. Image cubes at the top and label image at the bottom

3.4. Network architecture

(Ronneberger et al., 2015) introduced U-Net, a well-known deep-learning network for biomedical image segmentation. (Meng et al., 2023) has employed this network to segment the spine and vertebrae. Therefore, this project uses a 3D U-Net design. The U-Net architecture comprises the encoder (down-sampling) and the decoder (up-sampling). The encoder collects high-order abstract information from images while decreasing their size. At the same time, the decoder progressively recovers the original size of the input image. Additionally, skip connections between the encoder and decoder preserve information.

The U-Net is built with different blocks to facilitate code comprehension. Each block will be described in

the paragraphs below. The network takes several parameters as inputs, including the input channels and the number of feature maps at each level of the U-Net. The input tensor is initially passed through a convolutional block (Conv start) which performs the first feature extraction and encoding. This block encloses convolution operations followed by batch normalization and ReLU activation layers. Then, the resulting output is passed through four down-sampling blocks (Encoder) that consist of a max-pooling layer followed by a sequential process of repeated convolutional blocks. Here, the max-pooling reduces the spatial dimensions of the feature maps and extracts dominant features. These blocks gradually capture and encode hierarchical features at different scales.

After reaching the bottom of the U-shaped architecture, the tensor passed by a series of up-convolution blocks (Decoder). Each up-convolution block applies a sequence of convolutional blocks followed by an up-sample operation. Here, up-sampling acts as the reverse of the max-pooling, increasing the spatial dimensions of the feature maps and helping to recover the information lost during down-sampling.

An attention block is applied at each up-convolution block between the up-sampled features and the corresponding features from the skip connections. Inside every attention block, an attention mechanism combines the feature maps from the skip connections and the global feature maps to produce an attention map. The attention map is computed by applying convolutional operations to capture their interdependencies; this means that the attention procedure enhances the critical features and helps the model focus on relevant information. Therefore, the block output is a weighted combination of the input feature map based on the importance assigned by the attention system.

The up-sampled and attention-enhanced features are concatenated with the features from the skip connections. This concatenation combines the high-level and low-level features, allowing the model to retain and fuse information from different scales. Finally, the concatenated elements passed through the subsequent convolutional operations (up-convolution blocks) to further refine the features. The output of the final layer is routed through a 1x1x1 convolutional layer, and a sigmoid activation function is applied to generate the final result that provides the probability class at each voxel. The model used for U-Net with attention can be seen in the figure [Fig.7].

3.5. Training

Once the data was pre-processed, the network was trained. For the first training, which served as a reference, the following configuration was used:

1. the whole dataset without split
2. cube size of 96x96x96 with a stride of 48 and a voxel spacing of 0.37mm.
3. batch size of 10

With this configuration, it was noted that the training was prolonged. Due to GPU memory constraints, it was decided to modify the cube size to 64x64x64 with a stride of 32 and decrease the cube resolution, so the voxel spacing was increased to 1.0mm. Besides, the dataset was separated into three sets: train, validation, and test with 7, 2, and 2 CT scans, respectively, while the batch size remained the same. The training was repeated with these changes.

Other experiments were performed by changing the cube and stride size, raising and lowering the voxel spacing, and the batch size. However, the training was observed to be slower when the cube size was reduced, even though the batch size was increased.

Another thing that was considered was implementing a code with different blocks. One block for training, one for validation, and one for testing. This help to see the behavior of the network when training. Another parameter considered to vary according to the results obtained was the number of epochs, starting with 150 and 200 afterward. The latter is the number to be taken in the following training sessions. An instrument was also needed to visualize the plot´s behavior of the loss function, the accuracy, and the dice score in real-time. The TensorFlow tool: tensorboard was used for this purpose.

After several tests, the cube size chosen for all experiments was 64x64x64 with a stride of 32 and a voxel spacing of 1.0 mm. The appropriate configuration selected can be observed in the table [Tab.1].

Training parameters	
Loss functions	Dice and MSE
Optimizer	Adam
Learning rate	0.001
Batch size	56
Lambda(λ)	20
Epochs	200

Table 1: Optimal hyper-parameters configuration used in all the experiments after the first results.

Consequently, we proceeded to perform the training with the following variants:

- Training with optimal configuration
- Training with a weighted loss function
- Training with FOV adjustment with optimal configuration
- Training with FOv with a weighted loss function
- Group Bones training
- Group Bones training with a weighted loss function

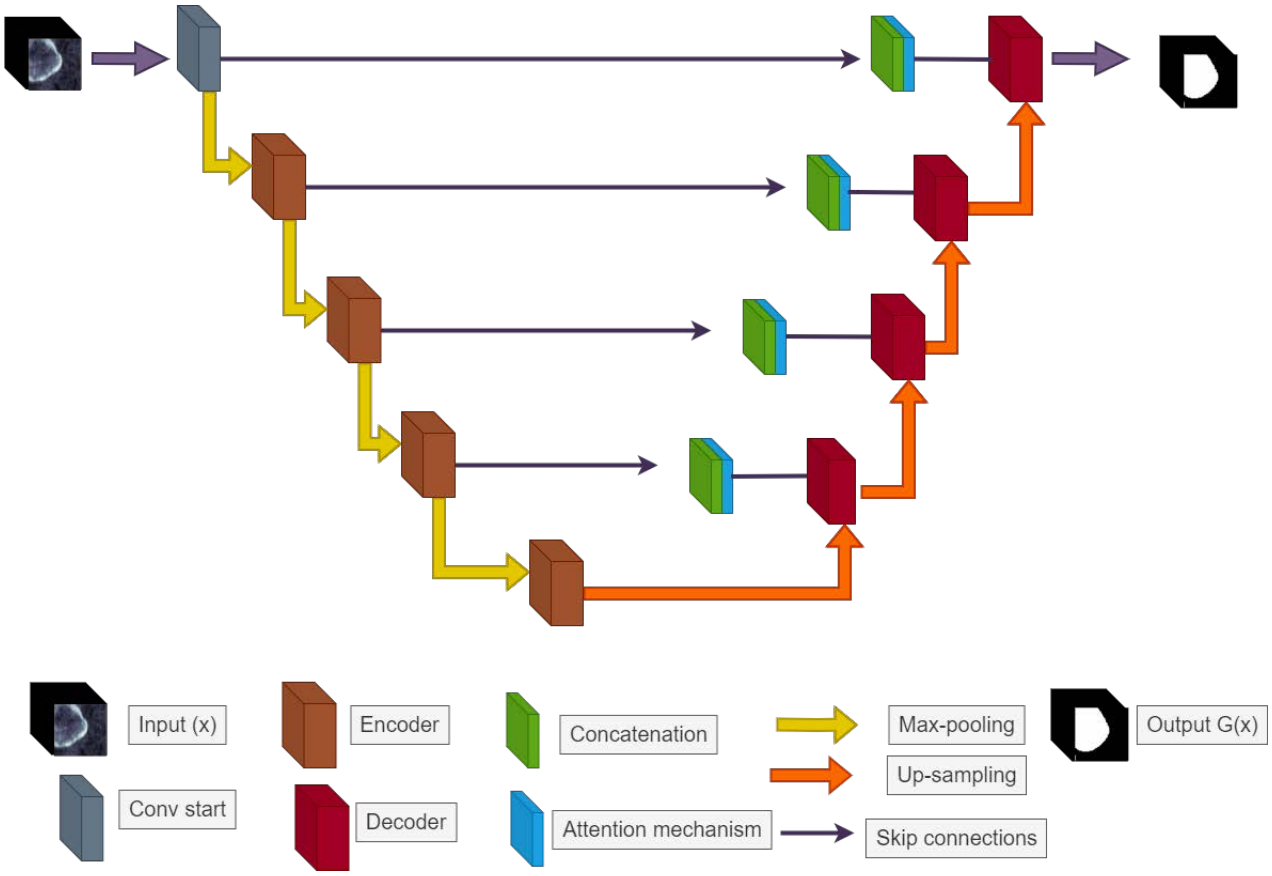


Figure 7: UNET 3D with attention architecture

3.5.1. Loss functions

In order to minimize the difference between our prediction and the actual data, the network was trained using a regression loss **L** (Meng et al., 2023). This loss function combines the Dice coefficient and the Mean Squared Error resulting in the following equation [Eq.1]:

$$\mathbf{L_{loss}}\,(G(x), y) = 1 - \frac{2|G(x) \cdot y|}{|G(x)|^2 + |y|^2} + \lambda \|y - G(x)\|^2 \quad (1)$$

Where the ground truth cube is denominated by y, the input by x, and the output by $G(x)$. λ is experimentally set to 20 and maintained constant during the training. The equation's first part consists of a dice loss function, and the second is an MSE loss function.

Dice Loss Function. For the dice loss, a function was created to calculate the dice coefficient, whose equation can be seen in [Eq.6]. In the dice coefficient, the best value is the one that is close to one; nevertheless, in a loss function, the values must be small since they are necessary to correct the weights in the back-propagation. Therefore, a modification is made to the equation used to calculate the dice by adding a one to subtract it. In this way, we obtain values close to the values that indicate that the model is learning and that it is good. The loss function is described by the following:

$$\mathbf{Dsc_{loss}}\,(G(x), y) = 1 - \frac{2|G(x) \cdot y|}{|G(x)|^2 + |y|^2} \quad (2)$$

MSE loss function. The MSE loss is obtained by the Pytorch library, and it measures the mean squared error between the predicted value and the target. The equation is shown in [Eq.3].

$$\mathbf{MSE_{loss}}\,(G(x), y) = \lambda \sum_{i=1}^{n} (y - G(x))^2 = \lambda \|y - G(x)\|^2 \quad (3)$$

3.5.2. Weighted Loss functions

One of our experiments involves training the model and implementing a weighted loss function for better segmentation results. A weighted loss function is a modified version of the standard loss function. The weights impose a higher penalty for incorrect minority class classifications. By raising the penalty of incorrectly classifying a minority class, the model will become more sensitive to that class.

In this case, we applied the weights to the bone class representing a small percentage in the image compared to the background class. A method was created to assign the weights to the loss function of each batch. It is implemented in the data loader and is computed with the following:

$$\mathbf{W_i} = \frac{max(S, \#bones_i)}{n} \tag{4}$$

Where S is a value selected according to the minimum numbers of non-zero voxels in the image, and it is used to avoid a result of zero. $\#bones$ is the number of voxels that contain bone, and n is the total number of voxels, which can be described by the size of the cube power of 3. Observing the number of bone voxels S is set to 6400.

3.6. Validation

A selection of the most optimal model is required to validate the trained experiments. Dice score (dsc) and accuracy are used for this purpose. Both metrics evaluate the performance of the segmentation task. Furthermore, the average loss function is also employed as a reference to determine which model is better than the other. The lower the average loss, the better the model. However, this selection is arbitrary because we must also take into account the values achieved in accuracy and dice in the elected epoch.

3.6.1. Metrics

In this case, a segmentation task's performance is evaluated. Two standard metrics are used.

Accuracy. The pixel accuracy is the percentage of correctly classified pixels in the image. It is computed with the following formula:

$$\textbf{Pixel accuracy} = \frac{\#TP + \#TN}{\#TN + \#TP + \#FP + \#FN} \tag{5}$$

In this case, the given classes are background (0) and bone (1):

- **True Positive (TP):** pixel classified correctly as bone
- **False Positive (FP):** pixel classified incorrectly as bone
- **True Negative (TN):** pixel classified correctly as background
- **False Negative (FN):** pixel classified incorrectly as background

However, there is an issue with this metric called class imbalance. If one class is more predominant than another, there may be a high accuracy indicating pixels are well classified while the other class is not. This result will be completely useless, and it demonstrates that high accuracy does not mean excellent segmentation skills. There are other metrics to deal with this problem. One of them is the dice score.

Dice score. The second measure is the Dice score which is defined as two times the intersection between the ground truth and the predicted mask, divided by the sum of the ground truth and the predicted mask. The formula is displayed in [Eq.6].

$$\textbf{Dsc} = \frac{2|GT \cap S|}{|GT|^2 + |S|^2} \tag{6}$$

The minimum value that the dice can take is 0, which means no intersection between the predicted mask and the ground truth, and the maximum that it can take is 1, which implies the prediction is accurate. Thus, the values closer to 1 indicate a decent overlap between the ground truth and the predicted mask.

3.7. Test

The test applied after the training & validation serves as an indicator of how good is the current model. A dice and accuracy score and an average of the loss to test the model's performance with unseen data were obtained. Nevertheless, the real test will be done on the whole CT scan, and it will be explained in the *Evaluation* 4.

3.8. Implementation details

This project was developed using an UBUNTU 22.04.1 LTS OS, Python 3.9.1, CUDA 11.2, and VSC 1.77.0 IDE. SimpleITK 2.2.1, 3D Slicer 5.2.1, ITK-SNAP 3.8.0, and MeshLab 2020.09 software were employed to help in image processing tasks.

In addition, the manipulation and reading of the NIFTI data were handled by the Nibabel 5.0.1, Nilearn 0.10.0, and Nipraxis 0.3.6 libraries. The deep-learning framework PyTorch 2.0.0 and its complementary packages were also utilized. The online TensorFlow tool, tensorboard, served to visualize the plots, time series, and model constitution. Lastly, the models were trained on two NVIDIA QUADRO RTX 5000 GPUs with 16 GB of built-in RAM each, provided by INRIA. With these specifications, each training session lasted approximately 16 to 20 hours.

4. Results

Two datasets to evaluate the model were used. One comes from INRIA as the training set, and the other from CHU Saint-Etienne. The first one has 2 CT subjects and two images per each, while the second one has 3 CT subjects and one scan per each.

The effectiveness of the segmentation is assessed using two common metrics from the literature (Payer et al., 2020). The first measure is the Dice score (DSC). It is calculated by dividing two times the intersection of the ground truth label and predicted label by the sum of the ground truth and prediction [Eq. 6]. The Hausdorff Distance (HD) is the largest of all the distances between a point in one set and the nearest point in the other set. It compares the predicted label with the ground truth.

On the one hand, the dice coefficient measures the segmentation accuracy and is particularly useful in dealing with imbalanced classes. In the case of these images, the background(the first class) is present in more regions than the bones(second class), giving; as a result, imbalanced classes, so there is a need for a metric to tackle this problem. The dice considering both the true and false positives gives a balanced calculation of the segmentation performance.

On the contrary, the Hausdorff distance reveals information about the boundary or shape dissimilarity between the masks. It is a helpful metric for evaluating how accurately the algorithm extracts the contours and fine details of the segmented objects. The foot contains various complex structures, and the segmentation entails clearly defining the outlines of the foot bones and separating them from the background or other entities. For this reason, if the segmentation is more similar, the Hausdorff distance is smaller; if it is more dissimilar, the HD is larger.

The first step for inference was to take the entire CT scan and apply the trained model in three different modes: overlapping, sliding, and central voting. These implement the same strategy used in cube extraction, obtaining, as a result, the binary mask.

Overlap. This approach divides the whole CT scan into cubes of size 64x64x64 as in the cube extraction to reproduce the same input as in training. Once the image is divided, the model is applied in an overlapping mode with a quarter of the cube size stride, in this case, 16. Each voxel contains 64 (4x4x4) probabilities, which are averaged and then binarized by setting a threshold of 0.5 to generate the final mask.

Slide. This technique splits the scan into cubes of 64x64x64 as in the overlap method but with the difference that the model is applied in a sliding mode. It involves moving a fixed-size window across the input image in a systematic manner to extract features. Likewise, the probabilities in each voxel are averaged and binarized to produce the mask.
It involves moving a fixed-size window or kernel across an input image in a systematic manner to perform local operations or extract features.

Central voting. In this mode, the image is divided into multiple cubes of size 64x64x64 with a stride of 16 (a quarter of the cube size), which is used to slide the cubes. Then, the model is applied to each cube by overlapping the 50%, and the other 50% of the cube is discarded, causing the central area to have more weight in the voting process. Here, each voxel has 8 (2x2x2) probabilities that are averaged and later, with a threshold of 0.5, are converted into binary values to get the mask.

Furthermore, the algorithm is programmed to save the best model as the one with the lowest validation average loss. Following this criterion, the best epoch is the one that contains the best model. However, when looking at the loss and accuracy plots, the lowest average loss is not the one that actually has the highest dice score value. Therefore the epoch that contains the highest average dice is selected. This procedure was applied to get the best model of all the experiments.

The results are exhibited in two ways: qualitatively and quantitatively. In the qualitative ones for the INRIA dataset, two types of images can be found. In the first image, the 2D slice representation, anterior and posterior foot, are displayed in three distinct views (axial, sagital and coronal) corresponding to different slices. Each window contains four images: at the upper left corner the GT, at the upper right corner the overlapping mask, at the bottom right the central voting mask and at the bottom left the sliding. This view allows a better comparison of all the masks obtained with the diverse methods. At the same time, the projections of the 3D meshes are also presented. Each window corresponds to the various modes employed for their creation; at the top is the central voting mesh, at the right is the sliding, and at the left is the overlapping.

The three modes for the CHU dataset are shown in the three views previously mentioned. Overlapping at the upper left corner, sliding at the upper right, and central voting at the bottom right. This dataset does not have Ground Truth, so it is not included.

In order to obtain this 3D perspective, the NIFTI images were subjected to a conversion algorithm to obtain volumetric structures. It examines each voxel and detects its relationship to the surrounding voxels. The intensities are then mapped to a color gradient and used to determine the iso-surface, representing the boundary between different forms. Besides, the 3D view allows a better analysis of the segmentation since here is possible to observe the missing or remnant parts in the mesh.

The images containing the 3D data are organized in such a way that the first model, from left to right, is obtained with central voting, the second with overlapping, and the third with sliding.

On the other hand, the quantitative ones show the metrics that indicate how good the results are numerically speaking. As the dataset of the CHU does not contain labels, the metrics could not be calculated, so

only qualitative results are presented.

This section will be divided to include all the specified experiments described in 3.5.

4.1. First experiments

The first experiment was used to evaluate the network's performance. The initial number of epochs was set to 150 and the batch size to 10 with a learning rate of 0.001 and using an Adam Optimizer, a lambda value of 20, and the loss function described in the 3.5.1.

4.1.1. INRIA dataset

The resulting masks are computed in the three modes mentioned above for both images of the two test subjects. The masks in 2D of the first results of the posterior and anterior foot are shown in the [Fig.9a]. The segmentation of the hind bones is much better than the anterior bones since a clear separation in the metatarsals cannot be distinguished.

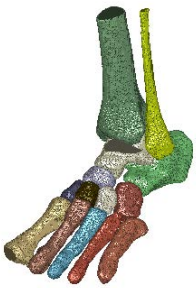


Figure 8: GT of the 3D Mask of the whole foot

Subsequently, the meshes of the whole foot are presented in [Fig.13a], and as a reference for how the 3D model should look, the labeled mesh in [Fig.8]. In this view, the foot structure is preserved but with some problems in the separations of the small bones. Also, it is possible to perceive the CT platform, which indicates that the images are not well-centered. These outcomes are taken as a reference to improve the current work.

On the contrary, good quantitative results are obtained for the first CT containing the posterior foot bones and being in the overlapping method, which presents the best dice score. However, in the other CT where the hind bones are also found, the Dice score decreases to 0.68, increasing the HD to 60.50 in the overlapping method and surrounding the same values in the sliding and the central voting. In contrast, for the scans containing the forefoot bones, such as the metatarsals, the dice remain in the range of 0.74-0.78 with Hausdorff distances with similar values. In the [Tab.2] are the metrics mentioned before for the three different techniques applied to the scans.

4.1.2. CHU dataset

The results achieved with this dataset are essential to determine how well the model generalizes with data it has never seen. The masks are obtained in the three different modes already mentioned. The masks in 2D of the first results are displayed in [Fig.11a]. Here, the whole structure of the foot can be appreciated. The metatarsus are seen in the axial and coronal views. The 3D meshes are shown in [Fig.14a]. This view matches well with what was seen in the 2D slice representation. Four of the five metatarsals are well-delineated, while the last is barely visible. The other bones have a good appearance but still need to be refined. Thus, with these results, it can be analyzed what should be improved in the subsequent experiments since complete modeling of the anterior zone of the foot is not yet achieved.

4.2. Training with optimal configuration

After observing the quantitative and qualitative results of the first training, the decision to train with more epochs was made. If the plots of average dice and average loss are observed well, it is possible to see that the average dice can still rise and the average loss can still go down. For this reason, the number of epochs was increased. The next training was executed, increasing the number of epochs to 200 and the batch size to 56. The rest of the hyper-parameters are shown in the [Tab.1].

4.2.1. INRIA dataset

The new experiment demonstrates a slight improvement in segmentation. In the figure [Fig.9b], it is possible to observe that some bones are almost separated than in the first training. Nevertheless, this disunion is more evident in the 3D masks [Fig.13b] since the metatarsals are more defined and with more separation between them than in the image of the first experiment. It means that the increase in epochs does influence obtaining better results. However, the same issue as before, the CT platform, is still present.

On the other hand, the metrics [Tab.3] showed that the results were slightly better in the first experiment than in this training. In addition, if a comparison is made between the different methods, it can be seen that the overlapping is the one that presents the highest dice scores and low HD surrounding values between 0.64-0.91 for dice score and 8.73-62.69 for the Hausdorff distances.

4.2.2. CHU dataset

In contrast to the first data set, a little deterioration is shown here. The metatarsals, well defined in the first experiment, are not so in this one, as shown in [Fig.11b]. However, the space between the first metatarsal and the phalanx present in the first experiment cannot be seen here; they are now cohesive. On the other hand, the larger bones can be seen better. This change can be observed in the meshes in [Fig.14b]. This image shows

	Overlap		Slide		Central votes	
ID	**Dice score**	**HD**	**Dice score**	**HD**	**Dice score**	**HD**
CT2 Posterior Foot	0.9221	8.4210	0.9027	14.0649	0.9177	8.2734
CT2 Anterior Foot	0.7811	10.6531	0.7627	18.7134	0.7597	16.0471
CT3 Posterior Foot	0.6834	60.5049	0.6627	62.1347	0.6767	59.3397
CT3 Anterior Foot	0.7590	21.7797	0.7562	24.0272	0.7438	22.1166

Table 2: Metrics for overlapping, sliding and central voting of the first experiments

	Overlap		Slide		Central votes	
ID	**Dice score**	**HD**	**Dice score**	**HD**	**Dice score**	**HD**
CT2 Posterior Foot	0.9165	8.7323	0.9041	26.5989	0.9107	14.6513
CT2 Anterior Foot	0.7229	22.7211	0.7671	23.9673	0.7055	23.6192
CT3 Posterior Foot	0.64789	62.6973	0.6256	63.0804	0.6445	61.5869
CT3 Anterior Foot	0.7640	22.3904	0.7645	25.9660	0.7402	22.5457

Table 3: Metrics for overlapping, sliding and central voting of the training with optimal configuration

that although the metatarsals are now missing segments, the larger bones, such as the fibula and talus, are more delineated.

4.3. Training with weighted loss function

The main objective of this experiment was to give more importance (weight) to the areas of the image with bone than the ones with the background. The segmentation masks provided by this model are presented below.

4.3.1. INRIA dataset

The results are better for the model trained with a weighted loss function. In the 2D view [Fig.9c], it can be appreciated that the bones on the bottom right corner in the image of the hindfoot are already separated. It is not observable in the previous results. The same happens with the bones of the anterior foot in the top right corner. Here, some bones look more divided than in the other experiments. In the 3D meshes [Fig.13c], it can be observed that the gaps between the metatarsals are more apparent than before, and the bones look more detailed.

Likewise, this could be proved with the metrics in the [Tab.4]. Those report an increment in the dice score of all of the scans except the first one and a diminution in the HD of all of the scans. Once again, the comparison between modes shows that the overlapping is the best due to its high dice scores and low HD.

The achieved enhancement is due to the weights added to the voxels with bones; this approach makes a difference in the appearance of the segmentation, even though the same problem is repeated since the CT table is still appearing.

4.3.2. CHU dataset

The masks produced with the model with weights can be appreciated in the [Fig.11c]. These demonstrate that some structures in the anterior foot reappeared again, like in the first experiment. However, in this trial, the more prominent objects are slightly disjointed, which is good because segmentation is becoming more accurate. Despite this, the metatarsus does not still look like the first experiment, but more so than the one with the optimal configuration. This difference is more visible in the 3D data,[Fig.14c], than in the 2D images. Additionally, the last metatarsal is a little more reconstructed here than before. This characteristic can be appreciated more in the overlapping mode (middle).

4.4. Training adjusting the image to the FOV with optimal configuration

The last results showed a problem with the masks. The CT table can be appreciated in the segmentation where the hindfoot bones are. Given this circumstance, it was decided to do a new training by adjusting the image to a suitable FOV. The qualitative and numerical results obtained with this experiment are shown below.

4.4.1. INRIA dataset

In this experiment, the results show a considerable improvement. It can be seen directly in the metrics [Tab.5], which indicate better results than previous ones. In the third scan, it is possible to see the increase in the dice and the noticeable decrease in HD. These values are due to the adjustment to the FOV of the image because, thanks to this, the CT table is no longer visible. Comparing the three modes, again overlapping demonstrates better scores than the others.

The new appearance can be verified with the images of the 2D [Fig.10a], which include bones and no other external factor as the platform. In the same way, the meshes in [Fig.13d] are cleaner than before, and now

	Overlap		Slide		Central votes	
ID	**Dice score**	**HD**	**Dice score**	**HD**	**Dice score**	**HD**
CT2 Posterior Foot	0.91344	7.7876	0.8999	19.8114	0.9083	11.9665
CT2 Anterior Foot	0.7760	18.4593	0.7369	20.3803	0.7574	18.8119
CT3 Posterior Foot	0.6696	62.2535	0.6737	62.9174	0.6606	60.7476
CT3 Anterior Foot	0.7970	16.7156	0.7500	18.4963	0.7746	23.9159

Table 4: Metrics for overlapping, sliding and central voting of the training with weighted loss

it is possible to observe better the whole 3D model. It should be highlighted that the characteristics obtained with the past training are preserved; the only difference is the disappearance of the table.

4.4.2. CHU dataset

In the case of this dataset, the table did not appear before, so this experiment is not significant at all. Nevertheless, it presents an enhancement in the details of the metatarsals than the segmentation seen in the previous experiment that can be observed in the 2D masks in [Fig.12a]. Similarly, the bones above the metatarsals show more delineated edges, indicating that the segmentation improves in slightly larger structures. For the 3D data,[Fig.14d], it is possible to see the better definition of the metatarsals but not the previously mentioned remarked borders.

4.5. Training adjusting the image to the FOV with weighted loss

This experiment is performed with the same FOV arrangement as before. The only difference here is that some weights are aggregated to the loss functions to increase the segmentation performance, give more importance to the bone voxels, and remove the external factors (the CT table) that caused some noise before. The results are displayed in the figures below.

4.5.1. INRIA dataset

When the weights are applied to the previous model, the dice score of all scans increases, lowering their Hausdorff distances. These metrics are presented in [Tab.6]. If their values are compared in the different modes, it can be seen that the overlap method is, another time, the best one. Despite the good metrics, the images do not reflect the same as they can be perceived in [Fig.10b]. For the posterior foot, the coronal view demonstrates that the previously separated bones are now a little closer together. On the other hand, the axial view of the anterior foot indicates that the metatarsals present more gaps between them. In contrast, the sagittal one suggests that these bones are under-segmented as they look like one.

In the meshes [Fig.13e] The only difference is that the metatarsals are farther apart, but neither improvement is particularly noticeable.

4.5.2. CHU dataset

For this dataset, the resulting masks are displayed in [Fig.12b] for the 2D and in [Fig.14e] for the 3D. In the 2D slice representation, the metatarsals are shown again, even more, complete than before. In the same way, the other bones are also more detailed. For the meshes, the bones look similar to the first results. Here, the metatarsals are delineated, but the phalanxes that were not present before can be barely perceived. It indicates that could be an over-segmentation in these masks because both datasets do not include the phalanxes, even though some images can have few remnants of them.

4.6. Training with separate structures

As the previous results showed some issues in specific structures, the decision was made to train four different groups of bone. The last experiment was divided into four groups of bones. They were selected accordingly to the shared features like function, size, and shape. The first group comprised the five metatarsals, the second group of the three cuneiforms, cuboid and navicular, the third one of the talus and the calcaneus, and the fourth of the tibia and fibula. These experiments were performed just on the first dataset as it is the one that can be compared qualitatively and quantitatively due to the labels. The different bones are displayed in the [Fig.15].

As seen in the meshes obtained from each group of bones, almost all present an adequate segmentation with some details to be improved. The best ones are those obtained in overlapping mode, as shown in the middle column of all the images and the metrics in the tables below. The only exception is found in group 2 of the cuneiforms since an adequate segmentation of these structures has yet to be achieved. Only a few bones can be observed in the sliding mode due to their shape, size, and organization. These bones are very close to each other, so the sliding way helps to reduce interference from neighboring bones by isolating each bone individually during the segmentation process.

On the other hand, two types of metrics are calculated. In [Tab.7] and in Appendix [Tabs.9,10 and 11] are shown the dice scores and HDs obtained between the GT and the segmentation mask only with a specific group of bones. While the [Tab.8] and in Appendix [Tabs.12,13 and 14] show a comparison between the

	Overlap		Slide		Central votes	
ID	**Dice score**	**HD**	**Dice score**	**HD**	**Dice score**	**HD**
CT2 Posterior Foot	0.9153	10.4259	0.9049	17.3032	0.9099	12.4377
CT2 Anterior Foot	0.7415	16.0471	0.7550	26.8142	0.7250	21.3096
CT3 Posterior Foot	0.9111	18.3738	0.8946	15.7761	0.9088	18.3439
CT3 Anterior Foot	0.7768	17.8254	0.7700	19.5925	0.7590	20.1404

Table 5: Metrics for overlapping, sliding and central voting of the training with FOV optimal configuration

	Overlap		Slide		Central votes	
ID	**Dice score**	**HD**	**Dice score**	**HD**	**Dice score**	**HD**
CT2 Posterior Foot	0.9170	5.2979	0.8964	17.3032	0.9127	7.9097
CT2 Anterior Foot	0.7732	12.7153	0.7805	16.6870	0.7472	13.0027
CT3 Posterior Foot	0.9095	17.9326	0.8973	17.5818	0.9070	17.9326
CT3 Anterior Foot	0.7987	22.9340	0.7558	21.3481	0.7846	23.3335

Table 6: Metrics for overlapping, sliding and central voting of the training with FOV with weighted loss

	Overlap		Slide		Central votes	
ID	**Dice score**	**HD**	**Dice score**	**HD**	**Dice score**	**HD**
CT2 Posterior Foot	0	75.7136	0	78.7309	0	78.3125
CT2 Anterior Foot	0.4299	39.2008	0.3098	39.2008	0.3638	39.2008
CT3 Posterior Foot	0	36.1030	0	76.8861	0	63.5916
CT3 Anterior Foot	0.4119	33.4252	0.2150	61.5447	0.4321	33.5376

Table 7: Metrics for overlapping, sliding and central voting of Group 1 bones

	Overlap		Slide		Central votes	
ID	**Dice score**	**HD**	**Dice score**	**HD**	**Dice score**	**HD**
CT2 Posterior Foot	0.7526	5.2846	0.2799	10.9447	0.7243	5.6113
CT2 Anterior Foot	0.9040	4.5012	0.8402	5.3362	0.8985	4.3149
CT3 Posterior Foot	0.7560	8.3229	0.7019	5.3490	0.7554	9.3090
CT3 Anterior Foot	0.8741	10.5304	0.8840	8.5100	0.8662	13.4682

Table 8: Metrics for overlapping, sliding and central voting of the Group 1 with all bones

GT focusing only on the specific group of bones and a mask containing all bones. In these results, it can be analyzed that the scores and HD are surrounding values above 0.75, reaching 0.90s. These measures are higher than the ones between the mask and the localized area; it means that the context given by the adjacent bones is essential to the model to delineate the objects' contours better. It is also worth mentioning that for the bones of groups 3 and 4, part of the hindfoot, no data are reported in the dice and HD as they do not exist in that perspective within the scan.

5. Discussion

In this section, the previous results will be analyzed. All of the experiments performed with the first dataset demonstrated that overlapping is the best method to segment. It can be noticed in all the 2D masks [Fig.9,10, 11 and 12]) and especially in the meshes created in this mode (middle column of [Fig.13,14 and 15]). As well as in the metrics calculated for the first dataset, since the highest dice scores and lowest Hausdorff distance were obtained with it. The models generated with this method present more defined and better-achieved structures than the central voting and sliding techniques. It happens because to produce the overlapping, 64 probabilities are collected, which are averaged and then binarized to determine if it is a bone voxel. Therefore,

this approach is more accurate because of the number of votes each voxel receives. On the other hand, central voting only receives eight votes that are also averaged and converted to binary values. Still, this method only focuses on the center of the voxel, discarding 50% of the information in the voxel. Thus, it also captures fewer features than in the overlapping mode. Likewise, the sliding receives only one vote, which leads it to collect less context and features that make it less robust when segmenting, and it is, in fact, the method that presents less definition in the qualitative results.

Furthermore, it is possible to observe that the bones that present better segmentation are the biggest ones. In the first dataset, the model performs well, segmenting the entire outline of the foot but not precisely on the metatarsal bones. These showed more complications at the time of segmentation. In contrast, the second dataset demonstrated that the metatarsals are also a complicated structure to segment, but these were well-defined in the experiment used as a baseline. Nevertheless, after all the experiments, the best shape was obtained in the model trained with the FOV adjustment and weighted loss.

On the other hand, the experiments done with the different groups of bones confirmed that the model accomplishes better results with large structures than with small ones. As can be observed in [Fig.15] and in the tables [Tabs.7,9,10 and 11], where the metrics of the group 3 which involves the talus and calcaneus are higher than the others. These bones are more prominent and more dense in comparison to the metatarsals. Moreover, the percentage of presence within the hindfoot scan is also above that in one foot containing the metatarsals. For this reason, these bones are challenging for the segmentation model.

In addition, group 2 of bones did not perform well during segmentation since these bones are barely appreciated in the scans of the first dataset. It is noticeable in the [Fig.5]. This group involves the three cuneiforms, the cuboid, and the navicular, particularly these bones have different and complex shapes, they are not disposed in the same way and are in proximity to other bones, so these facts complicate the segmentation process. Also, these bones have the same issue as the metatarsals; their presence in the image is less than other bones, as seen in [Fig.5], and are less dense than other bones.

5.1. Limitations

Despite the general good results, the model struggles to segment bones that share their borders with others, such as group 2, which is only partially present in masks. Likewise, small structures like the metatarsals are also hard to segment, leading to an under-segmentation or poor existence in the masks.

Indeed, the results result from low contrast in the scans and artifacts like the CT table, screws, or other biomechanical prostheses. Besides, not having enough data to train the network and the limited time for training caused GPU constraints that did not allow for added cross-validation and better time response to execute more experiments.

6. Conclusions

With all the results and models presented, the main objective of this project was achieved successfully. It consisted in performing a segmentation of the foot bones. However, this was completed in a global but not local context. That is to say that some bones present difficulties to be segmented, such as the metatarsals, cause of their size and presence in the scans. Along with the three cuneiform, the cuboid, and the navicular bones due to their intricate shape, small size, and different organization. Furthermore, the evaluation of both datasets demonstrates the ability of the model to detect all the bones, but only in part due to the challenging structures in the foot, despite the good metrics accomplished in some cases.

6.1. Future Work

The current results show that there is still work to improve segmentation and develop a fully autonomous method that relies only on the proper parameters and configuration to generate the desired results. Different techniques can be implemented, and some of them will be mentioned.

The first one is to train the network employing data augmentation methods since one of the limitations of the models is having few CTs to train and segment. In addition, performing cross-validation experiments could give better results since the model could be less biased.

The second thing to try is to modify the network architecture to be more robust. A UNETR could be helpful for this task because it can segment and identify the bones simultaneously without the need to divide the images into patches.

Furthermore, modeling ensembling can lead to better performance on the segmentation due to the variety and complexity between bones. One network can segment big bones, and the other can segment the small or problematic ones.

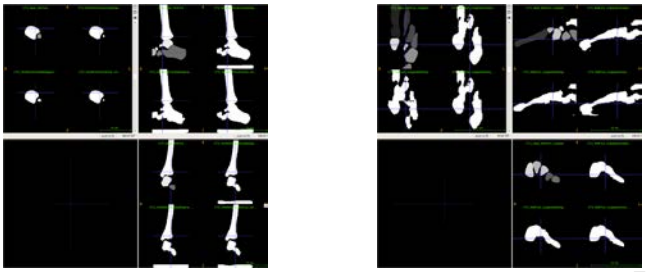
Acknowledgments

I want to thank all the INORA project partners for making this internship possible. Mainly to Centre Inria de l'Université Grenoble Alpes for providing the dataset and the training resources. And more importantly, to open its doors to complete my master's thesis, especially my supervisors Sergi Pujades and Julien Pansiot for the guidance and support I received at INRIA. In the

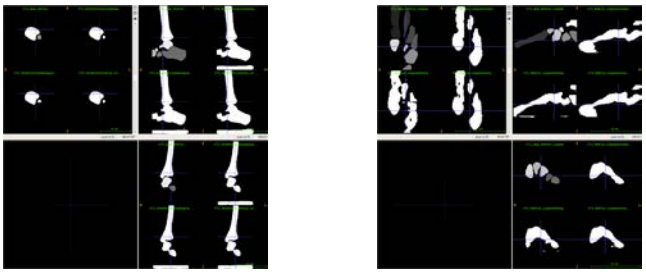
same way, I do my colleague Anne-Flore for motivating me every day and supporting me in difficult moments.

Furthermore, I am grateful to the MAIA program for selecting me and believing in my abilities to develop this master's degree. Also, to my coordinator Arnau Oliver, my professors, and the administrative team that make MAIA possible.

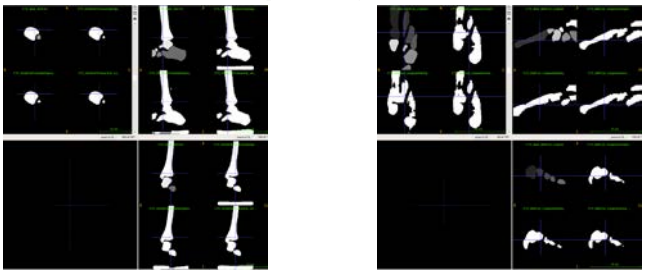
I would love to say that this master's is also dedicated to my mother and my family, who, although far away from here, have always motivated me with their words and given me the necessary strength to go ahead and finish this degree.



(a) First results

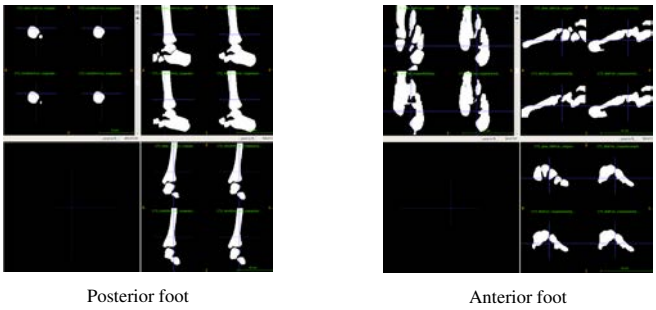


(b) Optimal configuration

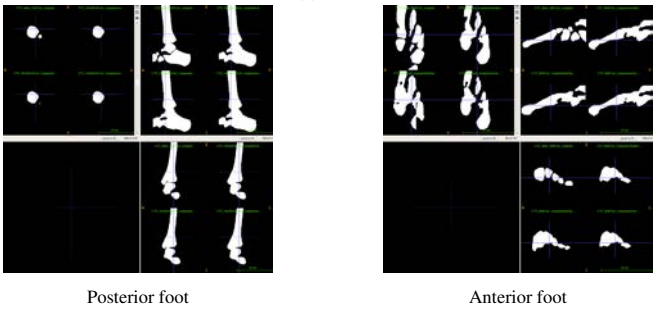


(c) Weighted Loss

Figure 9: 2D representation of INRIA dataset of Subject 1 (a) baseline experiment, (b) experiment with the best parameters, and (c) experiment adding weights to the loss function

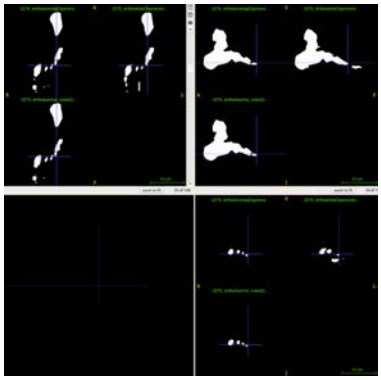


(a) FOV

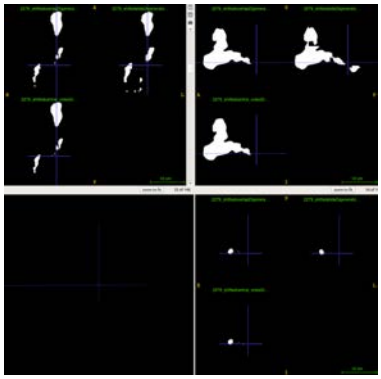


(b) FOV with Weighted Loss

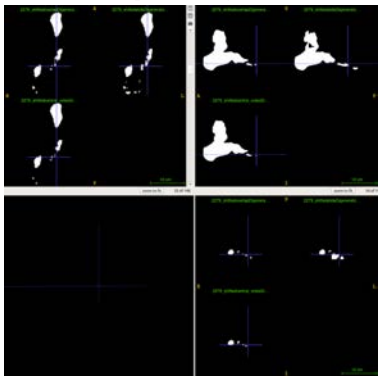
Figure 10: 2D representation of INRIA dataset of Subject 1 (a) experiment with the adjustment of the Field of View and (b) experiment with the adjustment of the Field of View adding weights to the loss function



(a) First results

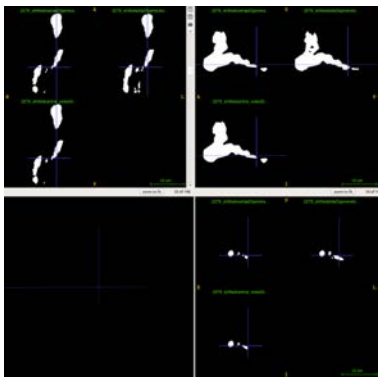


(b) Optimal configuration

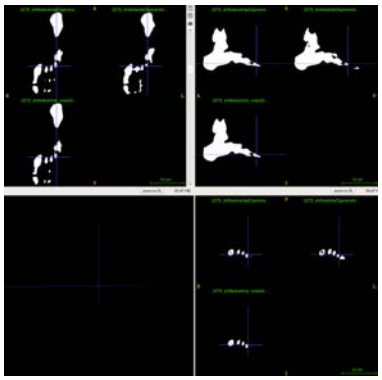


(c) Weighted Loss

Figure 11: 2D representation of St-Etienne dataset of Subject 1 (a) baseline experiment, (b) experiment with the best parameters, and (c) experiment adding weights to the loss function

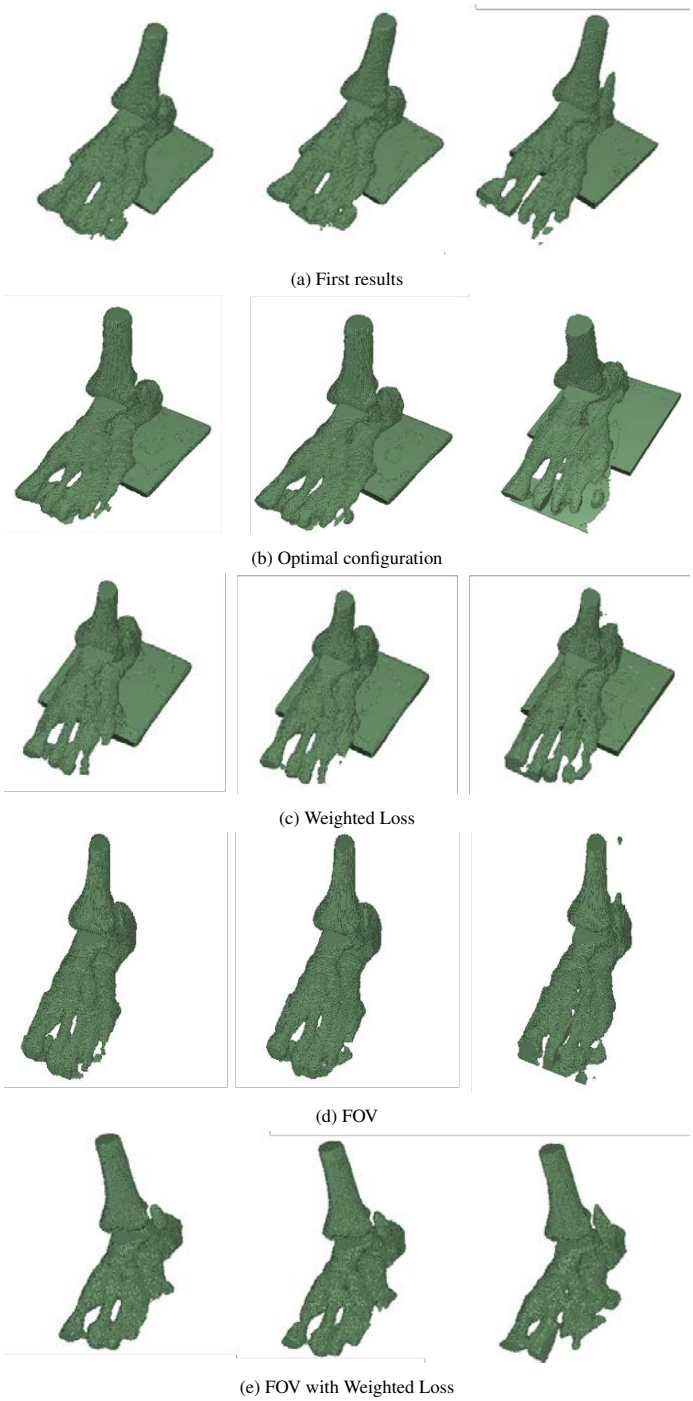


(a) FOV



(b) FOV with Weighted Loss

Figure 12: 2D representation of St-Etienne dataset of Subject 1 (a) experiment with the adjustment of the Field of View and (b) experiment with the adjustment of the Field of View adding weights to the loss function



(a) First results

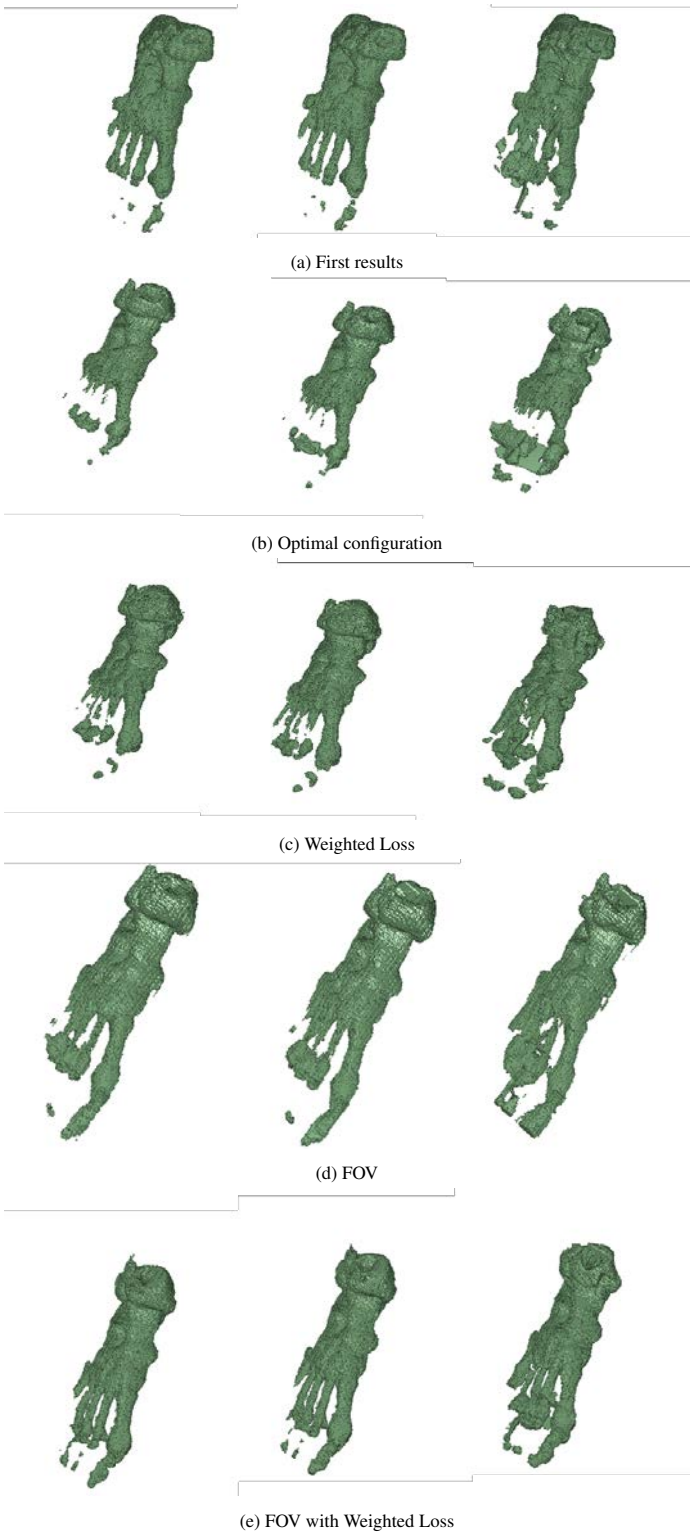
(b) Optimal configuration

(c) Weighted Loss

(d) FOV

(e) FOV with Weighted Loss

Figure 13: 3D meshes of the whole foot of INRIA dataset of Subject 1 (a) baseline experiment, (b) experiment with the best parameters,(c) experiment adding weights to the loss function, (d) experiment with the adjustment of the Field of View, and (e) experiment with the adjustment of the Field of View adding weights to the loss function



(a) First results

(b) Optimal configuration

(c) Weighted Loss

(d) FOV

(e) FOV with Weighted Loss

Figure 14: 3D meshes of the whole foot of St-Etienne dataset of Subject 1 (a) baseline experiment, (b) experiment with the best parameters,(c) experiment adding weights to the loss function, (d) experiment with the adjustment of the Field of View, and (e) experiment with the adjustment of the Field of View adding weights to the loss function

(a) Group 1: Metatarsals

(b) Group 2: Cuneiforms, cuboid and navicular

(c) Group 3: Talus and calcaneus

(d) Group 4: Tibia and Fibula

Figure 15: 3D meshes per group of bones of INRIA dataset of Subject 1 (a) Metatarsals, (b) Cuneiforms, navicular and cuboid,(c) Talus and calcaneus, and (d) Tibia and fibula

References

Aletaha, D., Smolen, J.S., 2018. Diagnosis and management of rheumatoid arthritis. JAMA 320, 1360. doi:10.1001/jama.2018.13103.

Chen, W.M., Lee, S.J., Lee, P.V., 2015. Plantar pressure relief under the metatarsal heads – therapeutic insole design using three-dimensional finite element model of the foot. Journal of Biomechanics 48, 659–665. doi:10.1016/j.jbiomech.2014.12.043.

Coughlin, M.J., Mann, R.A., 2014. Mann's surgery of the foot and ankle. Saunders Elsevier.

Egan, M., Brosseau, L., Farmer, M., Ouimet, M.A., Rees, S., Tugwell, P., Wells, G.A., 2001. Splints and orthosis for treating rheumatoid arthritis. Cochrane Database of Systematic Reviews 2010. doi:10.1002/14651858.cd004018.

Guillemin, F., Saraux, A., Guggenbuhl, P., Roux, C.H., Fardellone, P., Bihan, E.L., Cantagrel, A., Chary-Valckenaere, I., Euller-Ziegler, L., Flipo, R.M., et al., 2005. Prevalence of rheumatoid arthritis in france: 2001. Annals of the Rheumatic Diseases .

Hardy, M., Snaith, B., 2011. Musculoskeletal trauma: A guide to assessment and diagnosis. Churchill Livingston.

Kelly, E.S., Worsley, P.R., Bowen, C.J., Cherry, L.S., et al., 2021. Predicting forefoot-orthosis interactions in rheumatoid arthritis using computational modelling. Frontiers in Bioengineering and Biotechnology 9. doi:10.3389/fbioe.2021.803725.

Meng, D., Boyer, E., Pujades, S., 2023. Vertebrae localization, segmentation and identification using a graph optimization and an anatomic consistency cycle. Computerized Medical Imaging and Graphics 107, 102235. doi:10.1016/j.compmedimag.2023.102235.

Oktay, O., Schlemper, J., Folgoc, L., Lee, M., et al., 2018. Attention u-net:learning where to look for the pancreas. arXiv doi:10.48550/arxiv.1804.03999.

Parvizi, J., 2021. Orthopaedic knowledge update 13. Lippincott, Williams & Wilkins.

Payer, C., Štern, D., Bischof, H., Urschler, M., 2020. Coarse to fine vertebrae localization and segmentation with spatialconfiguration-net and u-net. Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications doi:10.5220/0008975201240133.

Pelt, M.N., 2012. Arthritis: Types, treatment, and prevention. Nova Biomedical.

Pham, D.L., Xu, C., Prince, J.L., 2000. Current methods in medical image segmentation. Annual Review of Biomedical Engineering 2, 315–337. doi:10.1146/annurev.bioeng.2.1.315.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. Lecture Notes in Computer Science doi:10.1007/978331924574428.

Standring, S., 2016. Gray's anatomy: The anatomical basis of clinical practice 41st ed. Elsevier.

Turner, D., Helliwell, P., Siegel, K.L., Woodburn, J., 2008. Biomechanics of the foot in rheumatoid arthritis: Identifying abnormal function and the factors associated with localised disease. Clinical Biomechanics 23. doi:10.1016/j.clinbiomech.2007.08.009.

Vergne-Salle, P., Pouplin, S., Trouvin, A.P., Bera-Louville, A., Soubrier, et al., 2020. The burden of pain in rheumatoid arthritis: Impact of disease activity and psychological factors. European Journal of Pain 24, 1979–1989. doi:10.1002/ejp.1651.

Wang, C.J., Hazlehurst, K.B., 2018. Orthopedic implant design and analysis: Potential of 3d/4d bioprinting. 3D and 4D Printing in Biomedical Applications doi:10.1002/9783527813704.ch16.

Zhang, M., Yu, J., Cong, Y., Wang, Y., Cheung, J., 2014. Foot model for investigating foot biomechanics and footwear design. Computational Biomechanics of the Musculoskeletal System doi:10.1201/b17439-3.

Appendix

	Overlap		Slide		Central votes	
ID	**Dice score**	**HD**	**Dice score**	**HD**	**Dice score**	**HD**
CT2 Posterior Foot	N/A	N/A	N/A	N/A	N/A	N/A
CT2 Anterior Foot	N/A	N/A	0	51.2820	0	52.3533
CT3 Posterior Foot	N/A	N/A	0	50.4870	N/A	N/A
CT3 Anterior Foot	0	50.9405	0.1766	31.4956	0	52.0754

Table 9: Metrics for overlapping, sliding and central voting of the Group 2 bones

	Overlap		Slide		Central votes	
ID	**Dice score**	**HD**	**Dice score**	**HD**	**Dice score**	**HD**
CT2 Posterior Foot	0.9202	6.7922	0.4957	34.0400	0.9219	7.5917
CT2 Anterior Foot	N/A	N/A	N/A	N/A	N/A	N/A
CT3 Posterior Foot	0.7557	24.2371	0.4901	36.9611	0.7572	22.4849
CT3 Anterior Foot	N/A	N/A	0.1952	32.2940	0.1842	36.6841

Table 10: Metrics for overlapping, sliding and central voting of the Group 3 bones

	Overlap		Slide		Central votes	
ID	**Dice score**	**HD**	**Dice score**	**HD**	**Dice score**	**HD**
CT2 Posterior Foot	0.2836	41.9146	0.2528	42.2675	0.2871	41.9146
CT2 Anterior Foot	N/A	N/A	N/A	N/A	N/A	N/A
CT3 Posterior Foot	0.3030	40.8628	0.5347	33.2814	0.3287	40.8628
CT3 Anterior Foot	N/A	N/A	N/A	N/A	N/A	N/A

Table 11: Metrics for overlapping, sliding and central voting of the Group 4 bones

	Overlap		Slide		Central votes	
ID	**Dice score**	**HD**	**Dice score**	**HD**	**Dice score**	**HD**
CT2 Posterior Foot	0.9299	8.7792	0.8624	8.5180	0.9253	8.5421
CT2 Anterior Foot	0.9702	7.8051	0.9623	7.8051	0.9734	7.7700
CT3 Posterior Foot	0.9202	8.3966	0.8872	8.3147	0.9089	9.1083
CT3 Anterior Foot	0.9628	6.9811	0.9384	7.5556	0.9647	7.6635

Table 12: Metrics for overlapping, sliding and central voting of Group 2 with all bones

	Overlap		Slide		Central votes	
ID	**Dice score**	**HD**	**Dice score**	**HD**	**Dice score**	**HD**
CT2 Posterior Foot	0.9716	3.7732	0.9537	7.0688	0.9724	3.7732
CT2 Anterior Foot	0.9176	5.7914	0.8937	6.8324	0.9199	5.6113
CT3 Posterior Foot	0.9667	4.5616	0.9535	6.3008	0.9663	4.5616
CT3 Anterior Foot	0.8661	10.6338	0.8698	7.9786	0.8777	10.4258

Table 13: Metrics for overlapping, sliding and central voting of Group 3 with all bones

	Overlap		Slide		Central votes	
ID	**Dice score**	**HD**	**Dice score**	**HD**	**Dice score**	**HD**
CT2 Posterior Foot	0.9482	4.0700	0.9545	3.8805	0.9469	4.2186
CT2 Anterior Foot	N/A	N/A	N/A	N/A	N/A	N/A
CT3 Posterior Foot	0.9034	18.3737	0.8940	15.7760	0.9035	18.3737
CT3 Anterior Foot	N/A	N/A	N/A	N/A	N/A	N/A

Table 14: Metrics for overlapping, sliding and central voting of Group 4 with all bones

Medical Imaging and Applications

Master Thesis, June 2023

Deep Learning Explainability for Breast Cancer Detection in Mammography

Karla Guadalupe Sam Millan, Prof. Robert Martí

ViCOROB, Universitat de Girona

Abstract

Deep Learning's recent popularity in the field of Medical Imaging and CADx has generated concerns over the need to understand the decision-making process of the models, which has caused the development of different explainability methods (XAI). This study applied several of these XAI algorithms, namely saliency maps, Occlusion, Integrated Gradients, Guided GradCAM, LIME, SHAP, and DeepLIFT, to evaluate the performance of two CNNs trained on two different classification tasks: patch-based breast cancer classification and whole mammogram classification. The attribution maps obtained from the first task were qualitatively evaluated, and it was found that true positive predictions tended to highlight the lesion's area, while true negative predictions had more spread out highlighted regions. The attribution maps from the second task showed that the CNN highlighted the area of the mammogram where the lesion was located. The lesions were also highlighted in false negative and true positive mammograms with low probability scores, which implies that the model underwent correct training and learned the relevant features. Considering the idea that there should be a connection between explainability and the position of the lesions, IOU scores were computed to quantitatively evaluate the different XAI approaches. Integrated Gradients performed best at locating the lesions, while SHAP and LIME were the worst-performing.

Keywords: Deep Learning, explainability, XAI, mammography, attribution maps, breast cancer, classification, occlusion, SHAP, saliency maps, integrated gradients, DeepLIFT, Guided GradCAM

1. Introduction

Deep Learning has risen in popularity in recent years in the field of Computer Aided Diagnosis (CAD), with current state of the art implementing different deep learning models for solving a wide-range of tasks in medical imaging (Singh et al., 2020). One such task is breast cancer classification. As one of the leading causes of mortality amongst women, breast cancer has spurred significant interest in developing improved detection methodologies. Due to the vast amount of breast cancer cases, CAD systems seek to alleviate radiologists' workload, aiming to reduce work hours and improve detection accuracy (Balkenende et al., 2022). Convolutional Neural Networks (CNNs), in particular, have been extensively used for the purpose of classifying mammographies as benign or malignant (Loizidou et al., 2023). Arevalo et al. (2015) were among the first to use CNNs for classifying breast lesions, achieving an AUC value of 0.86 on the BCDR-F03 dataset. Since then, various other CNN architectures have been applied, like ResNet-50 and InceptionResNet-V2, where the latter achieved a 97.5% and 95.3% accuracy on the DDSM and INbreast datasets respectively in a study by Al-antari et al. (2020). Research findings have demonstrated that these CNN-models exhibit comparable performance to that of radiologists, and in certain instances, can increase the proficiency of radiologists when employed as a supplementary tool (Ou et al., 2021). Although DL methods have continuously been proven to perform adequately as CAD methods for Breast Cancer, there have been growing concerns as to the use of this technology. Deep Learning models are essentially considered as "black boxes", due to the sheer number of layers and weights inside, rendering it practically impossible to completely understand the inner workings and mechanisms of the neural network. This is critical in the medical field, where a decision made based on the results obtained from a neural network could have a huge direct impact on a patient's life. Moreover,

DL should comply with regulations like the European Union's General Data Protection Regulation (GDPR) (van der Velden et al., 2022). Explainable AI (XAI) aims to alleviate these previously stated concerns by developing various techniques that seek to understand deep learning algorithms. Some of these XAI methods include Saliency Maps, Integrated Gradients, Occlusion, Guided GradCAM, LIME, SHAP and DeepLIFT.

The objective of this work is to visualize and evaluate the performance of a trained classifier for 2D mammographies with the different XAI algorithms , and to compare the differences between them. The study was performed as follows:

1. Patch based XAI: To train standard CNNs on mamography with normal and malign patches, and assess the results of the different XAI methods on the model.
2. Whole mammogram XAI: To train models on full mammography images to classify between malignant and healthy images and assess the results with several XAI methods.
3. Evaluation of XAI methods: Qualitative and quantitative evaluation of the XAI attribution maps in terms of robustness and localization of findings with bounding boxes and IOU scores.

2. State of the art

2.0.1. XAI Methods

Saliency maps are considered as the baseline approach in XAI for medical imaging, and provide insights to the parts of an image that contributed the most towards a prediction from a neural network (van der Velden et al., 2022). They work by computing the gradients of the target class score with respect to the input. The output is a matrix of similar shape to the input image, where values close to 0 correspond to pixels which have a smaller impact on the output. High values, either positive or negative, on the other hand, correspond to pixels which majorly affect the output score (Simonyan et al., 2014).

Guided GradCAM is a point-wise multiplication between Guided Backpropagation and GradCAM, which makes it able to function with any type of CNN. It works by computing the gradient of the score for a given class with respect to a feature map, and it then applies global average pooling. Then, it obtains a weighted combination of the feature maps, and a RELU is subsequently applied to only acquire the features which influence positively the result. The resulting heatmap has the same size as the feature maps, and thus needs to be resized (Selvaraju et al., 2019).

Integrated Gradients compute the gradients of the output from the model for each step along a linear path between a baseline, which can be a blank image in the case of images, and the input. The gradients are then integrated, which accumulates the changes as the input goes from the baseline to the input image. A characteristic of this method is that it doesn't need to modify the network (Sundararajan et al., 2017).

Occlusion perturbs the image by applying a gray sliding window along the image, noting the changes in the output for each window position. Thus, when the correct class is occluded by the window, the output probability is expected to drop significantly (Zeiler and Fergus, 2013).

Like Occlusion, LIME (Local Interpretable Model Agnostic Explanations) perturbs the image and computes its corresponding output score for each of the perturbed images. This perturbation occurs at a pixel or superpixel level. The latter is done by providing LIME with a segmentation mask that divides the image into different regions. Then, a simpler interpretable model is trained on the perturbed images and their predictions, in order to learn the relationship between the changes to the original image and the original model's predictions. The higher the weights assigned to a pixel or superpixel, the higher its impact on the prediction (Ribeiro et al., 2016).

Based on cooperative game theory, SHapley Additive exPlanations (SHAP) uses Shapley values to compute the magnitude of the contribution of each feature to the output of the model. It achieves this by perturbing the image and altering the pixels (or superpixels). To compute the contributions, SHAP considers the different subsets of features and calculates the output with each given subset until it considers all possible combinations. However, this is computationally intensive, which has spawned some variations to approximate Shapley Values, such as Kernel SHAP and Deep SHAP (van der Velden et al., 2022)

DeepLIFT addresses the saturation problem by introducing a reference input and its reference activations in the network. The reference activations are compared to new activations to compute the activation changes and, therefore, the contributions of each neuron (de Vries et al., 2023) (Shrikumar et al., 2019).

The following section will list some of the state of the art applications of XAI in Mammographies.

2.0.2. XAI in Mammography

XAI methods have been used in the medical imaging field used to alleviate the concerns surrounding the black-box paradigm, as previously stated. Several visual explainability approaches have been used to study and understand a wide range of image modalities from various anatomical locations, like: brain, breast, cardiovascular, chest, prostate, eye, skin, among others (van der Velden et al., 2022). In the case of breast applications, researchers have obtained visual explanations for X-rays, MRI, ultrasound, and histological images.

Regarding XAI in 2D mammographies specifically, XAI has been used to evaluate the performance of the

networks on various problems. Huang et al. (2020) proposed a hybrid neural network comprised by two parts: a modified PCANet and a DenseNet. They compared their proposed HybridNet with other popular models like PCANet, ResNet and DenseNet, and found that HybridNet outperformed all of them. To confirm its correct performance, CAM was applied, and it was observed that the resulting attribution maps focused on the abnormal parts (i.e. the lesions) of the mammographies, indicating that HybridNet had successfully learned the important features for the classification problem.

Akserlod-Ballin et al. (2019) developed an ML-DL model that combined mammographies and clinical data to detect breast cancer, resulting in an AUC of 0.91, a specificity of 77.3% and a sensitivity of 87%. The combination with clinical data provided a level of interpretability to the model, and this was further explored by computing the impact of each of the data's features via SHAP.

Xi et al. (2020) tackled the problem of having high-resolution mammographies with meaningful information located in very small regions of the image. Instead of resizing the full images, which entails a loss of information, they trained several CNNs like AlexNet, VGGNet, GoogLeNet, and ResNet with the cropped ROIs, and and then applied created abnormality detectors by integrating the CNNs with either CAM or other region proposal networks. ResNet was the network selected for integration with CAM, and found that the detection results obtained from the heatmap aligned with the ground truth for the lesion localizations.

A paper by Kobayashi et al. (2022) utilized generative contribution mapping (GCM). GCM is a classification model proposed by Arai and Nagao (2017) that uses XAI to explain its classification predictions by creating class contribution maps and class weight maps. Kobayashi et al. used this method to classify the existence of calcifications on mammographies. They found that GCM was more efficiently explainable when combined with class contribution maps and class weight maps than with GradCAM. It also found that GCM, when used with the maps, could provide important visual information even in the case of a false negative, due to the highlighting of the microcalcification localizations even with an incorrect prediction.

Yi et al. (2019) aimed to develop a CNN to classify mammography images according the view, the breast laterality and the breast density. The model architecture selected for the three tasks was ResNet-50, modifying its last layer and assigning it either two or four output neurons, according to the task. CAM was applied to visualize the network's decision when predicting for each of the three objectives. The model showed an AUC of 1 for mammographic view and 0.93 laterality classification, but it displayed a 68% accuracy when classifying breast tissue density. CAM's heatmaps displayed the network's focus on the superior part of the image generally corresponding to the pectoral muscle for the mammographic view task. As for the laterality task, the heatmap highlighted the region located towards the left or right, depending on the laterality. Even though the third task achieved such a low accuracy, the heatmaps consistently highlighted the regions corresponding to the breast, even if the network predicted an incorrect breast density, indicating that it correctly based its decision on the breast tissue.

Prodan et al. (2023) developed both CNN and Visual Transformers for breast cancer detection. They employed a data augmentation technique involving synthetic images to reach better performance. After training the models, they made use of GradCAM and bounding boxes to gain insight into their models' decision making procedure and behavior.

3. Material and methods

3.1. Dataset

The dataset used for the breast patches classification task is the Iceberg Selection, a subset of the OPTIMAM Mammography Image Database (OMI-DB) (Halling-Brown et al., 2020), consisting of patches centered on breast lesions and patches with normal breast tissue. There are a total of 3808 full-images acquired from three different scanners: Hologic, Siemens, and GE, with each image having a patch centered around the lesion, and a normal patch, thus yielding a total number of 7616 image patches. The dataset was divided into training (80%) and validation (20%) subsets, ensuring no overlap of patients between them to avoid bias.

For the full-mammography classification task, two datasets were used. The first dataset is a more balanced subset of the training data of the RSNA22 challenge (Carr et al., 2022). It has a total of 2767 images, of which 1647 were negative cases, and 562 were positive (malignant) cases. The images in the dataset were all preprocessed, flipping the images to have the same laterality (left), cropping the background, and saving them as PNG files. Like in the previous task, the dataset was divided into 80% training and 20% validation, avoiding patient overlap. The second dataset contained the malignant full-images from the OMI-DB subset described previously that were acquired with the Hologic scanner, due to the higher similarity of the images from the RSNA22 dataset in comparison to those acquired with the Siemens or GE scanners. As for the benign images, an equal amount as the malignant images were selected from the OMI-DB dataset, which were also subsequently flipped, cropped and saved as PNG files. Thus, the total number was 7229 full-images. Training and validation were divided 80%:20% with no patient overlap. Table 1 summarizes the datasets used.

		Train	Validation	Total
OMIDB Subset (Iceberg Selection)	**Non-malignant**	3045	763	3808
	Malignant	3045	763	3808
	Total	6090	1526	**7616**
RSNA22 Subset	**Non-malignant**	1647	411	2058
	Malignant	562	147	709
	Total	2209	558	**2767**
OMIDB Hologic Subset (Full-Images)	**Non-malignant**	2896	719	3615
	Malignant	2892	722	3614
	Total	5788	1441	**7229**

Table 1: Datasets used for the breast classification tasks.

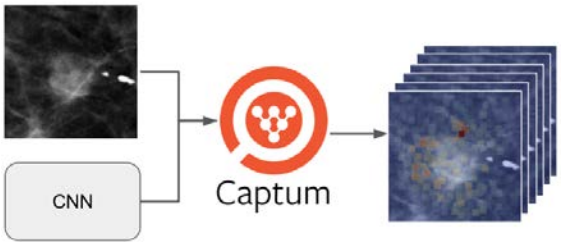


Figure 1: General diagram showing the generation of the heatmaps with Captum. Each of the explainability methods takes the CNN (either the ResNet-50 or EfficientNetB0) and an image (patch or whole mammogram) as inputs, and generates an attribution map containing the contribution scores for each pixel or superpixel.

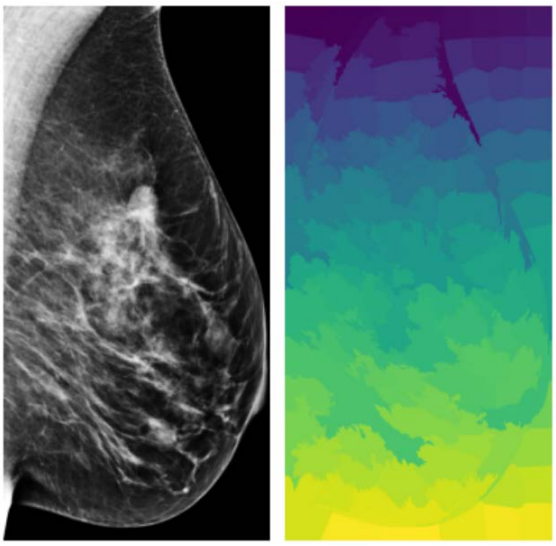


Figure 2: An example of segmentation performed with Scikit-Image's SLIC. The segmented image is fed as a feature mask for the generation of SHAP and LIME's attribution maps.

3.2. Methods

3.2.1. Preprocessing

The whole mammograms in from the Iceberg Selection were previously flipped, cropped and saved as PNG files. In order to create the Normal + Malignant OMI-DB Hologic database, the normal images were preprocessed in the same way. The DICOM files were read, thresholded, and a bounding box was computed with OpenCV's ConnectedComponents and findContours. The images were then cropped according to the computed bounding box to remove the background. Finally, the mammograms were saved as PNG files.

3.2.2. Breast Patches Classification

MobileNetV2 and ResNet-50 are two popular CNN architectures, and were therefore selected for this problem. Both were trained on the Iceberg Selection Dataset, with the train set image transformations consisting of a horizontal flip, a vertical flip, and a random rotation of 30 degrees for data augmentation, and a 224x224 resizing and normalization. The validation set transformations consisted only of the 224x224 resizing and normalization. The loss used was Cross Entropy Loss, the optimizer was Adam with a learning rate of 0.001, and a ReduceLROnPlateau scheduler with a patience equal to 5.

Out of both of them, ResNet-50 performed the best, with an accuracy of 97% compared to MobileNetV2's 93%. ResNet-50's high performance was in line with what was expected, as the classification problem was relatively simple due to the very different appearances of a normal patch vs one with a malignant lesion.

3.2.3. Full-Image Classification

Due to ResNet-50's good performance in the previous problem, several experiments were initially performed with it for Full-Image Classification with the RSNA22 subset dataset. Since it was an unbalanced problem with around three times the amount of negative cases vs positive cases, the weight of the positive examples was set to 3. Several attempts with varying learning rate values were made. However, ResNet-50 failed to yield satisfactory results.

The next network attempted was EfficientNetB0. Cantone et al. used this architecture, among others, for classifying whole mammograms. They used SGD with a momentum of 0.9 and a Cosine Annealing scheduler with warm restart, and varied the input image resolution Cantone et al. (2023). Thus, the same hyperparameters were tested in this study, in addition to the focal loss. The selected learning rate was set to 0.01, and the input size was set to 1024x512, as higher resolutions did not perform significantly better and its computational costs were substantially higher. This approach with the RSNA22 subset reached an AUC of 0.72.

Aiming to further improve the performance of the network, EfficientNet-B0 was then trained on the Normal + Malignant OMI-DB Hologic subset previously described. The hyperparameters selected were Cross Entropy Loss, SGD with momentum equal to 0.9 and a learning rate of 0.01, and Cosine Annealing with warm restart as a scheduler. The input size was kept at 1024x512. The AUC reached was higher than in the previous step, with a value of 0.83.

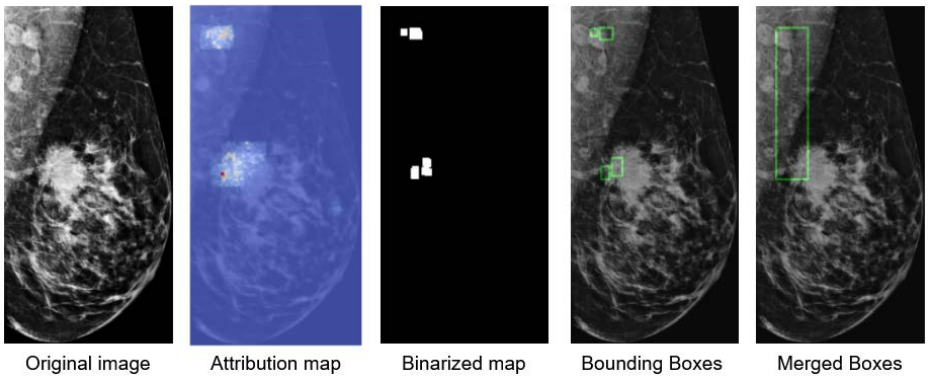


Figure 3: Steps to acquire the bounding boxes from the attribution maps. The attribution maps are obtained with a XAI technique, and they are subsequently binarized by keeping only the pixels with a score above a selected quantile, which are then cleaned by keeping only the larger regions via morphological operations. The bounding boxes are then acquired with the contours of the binarized maps, and they are finally merged into a large bounding box.

3.2.4. Explainability

Once the ResNet-50 and EfficientNetB0 models were trained and selected for each of the tasks, several explainability methods were applied on them to observe the inner workings and performance of the models. Captum for PyTorch is an open-source library that includes many explainability methods (Kokhlikyan et al., 2020). As such, the following attribution methods were computed utilizing this library. Fig. 1 summarizes the generation of the attribution maps by evaluating the model's contributions with an input image.

Computing the vanilla saliency maps and the DeepLIFT attribution maps require solely the trained model and the target class for which the gradients are computed. The attribution maps obtained initially with both of these methods are hard to visualize, given the large size of the input image and the small highlighted regions. To solve this, the attribution maps were dilated with a rectangular kernel of size 9x9 in a single iteration.

Integrated gradients, like the previous two, require only the target class. The number of steps performed by the approximation method for the integrals was set to 200. This was selected as a compromise between computational cost and attribution map resolution. As with vanilla saliency and DeepLIFT, the acquired maps were dilated for better viewing.

Guided GradCAM requires the specific layer for which the attributions are to be computed. The last layers for both models were specified and the attribution maps were also dilated.

As a perturbation-based approach, Occlusion required the shape of the patch with which to occlude the input image, and, optionally, the strides that the patch should take in each direction after every iteration. The sliding window shape was set to (3, 60, 60) and the strides to (3, 30, 30). The relatively large window size and strides were a compromise for the large computational times and the resolution of the attribution maps.

Captum's LIME approach takes in an optional feature mask argument, which groups the image's pixels into superpixels, and treats each group as a single interpretable feature. If a feature mask is not provided, then LIME considers each pixel as an individual interpretable feature, which largely increases their number, resulting in very slow attribution map computations. Thus, the feature mask is obtained by dividing the input images into 150 segments using Scikit-Image's SLIC method. Fig. 2 shows an example of a mammogram segmented with SLIC. Afterwards, the feature mask is fed into LIME with a number of steps equal to 200. The resulting attribution maps were not dilated, as the superpixels are easily observable. SHAP, much like LIME, was fed the same feature mask, on account of the same reasons. The attributions were not dilated either because of the superpixel grouping. An example of a feature mask obtained with SLIC can be observed in Fig.2.

3.2.5. Quantitative Evaluation: IOU

Using OMI-DB's subset that solely includes images with lesions, it was hypothesized that the explainability results should show a certain relationship with the position of the lesions. To this effect, and to evaluate the different explainability methods' attributions in the whole mammogram classification task, the Intersection Over Union (IOU) was calculated with respect to the ground truth bounding boxes available for the malignant full-images from the OMI-DB subset. Bounding boxes generated from the attribution maps were therefore needed. To achieve this, the following steps were taken:

1. The dilated attribution maps were selected (except for LIME and SHAP) because they generated larger connected regions corresponding to the mass' location which were not as affected by the morphological operations from the next steps. These dilated attribution maps were binarized by thresholding them according to a given quantile. If

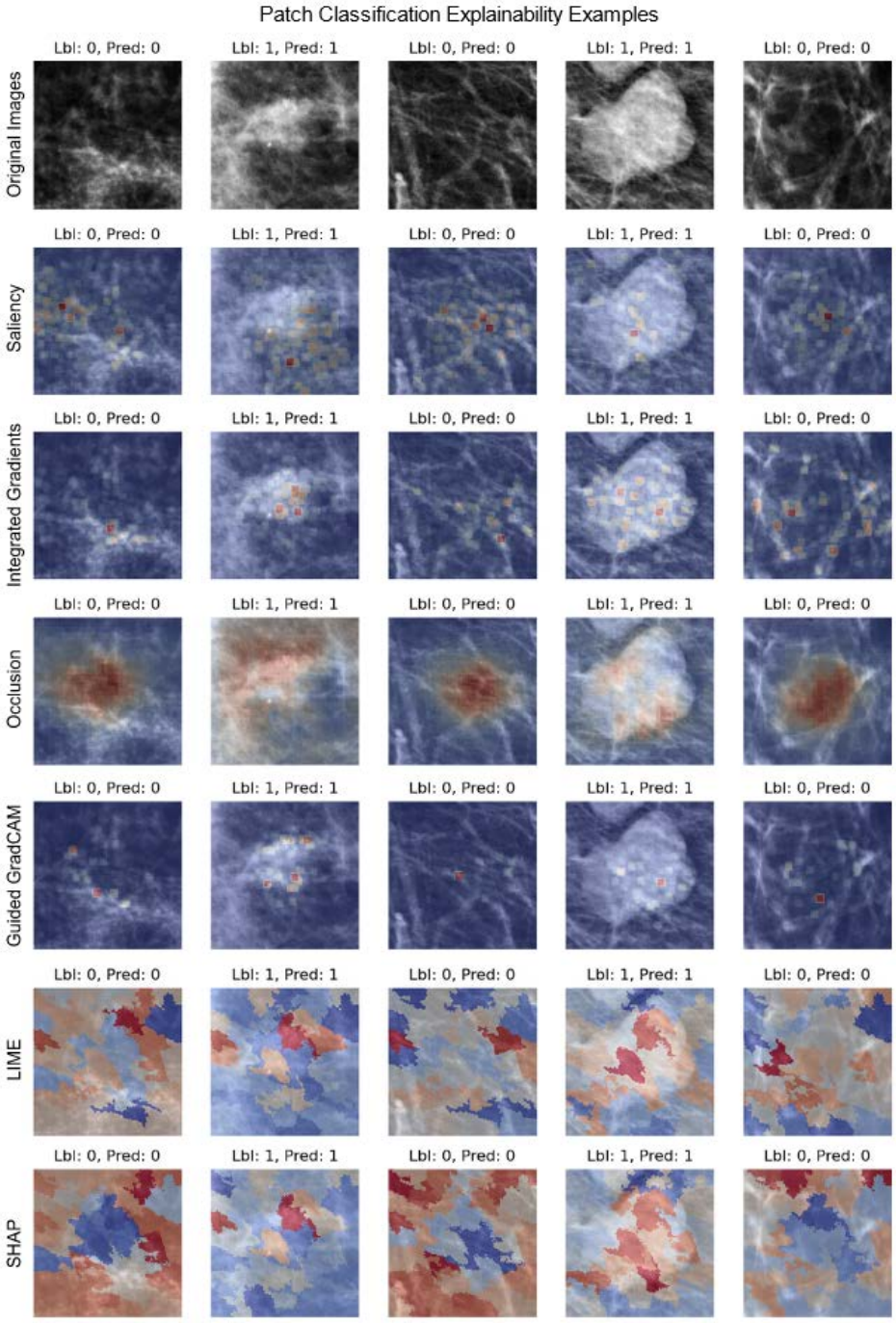


Figure 4: Examples of mammography patches that are either non-malignant or malignant. The top row corresponds to the original images, and the following rows show the attribution maps generated with a different explainability technique. Red areas correspond to higher attribution scores and higher impact on the model's prediction, while blue areas represent low scores.

the attribution score for a pixel was higher than that quantile, the score would be set to 1. Otherwise, it was set to 0.

2. The binarized attributions were eroded and then dilated, to remove the very small regions.
3. The bounding boxes for each separate region were obtained by finding their contours and generating a box for each contour.
4. In the case of multiple bounding boxes in a single image, they were combined into a single big bounding box with the minimum and maximum xy coordinate values found among the multiple boxes and setting them as the upper left and lower right coordinates for the combined bounding box, respectively.
5. Finally, the IOU scores for each image for each of the explainability methods were computed.

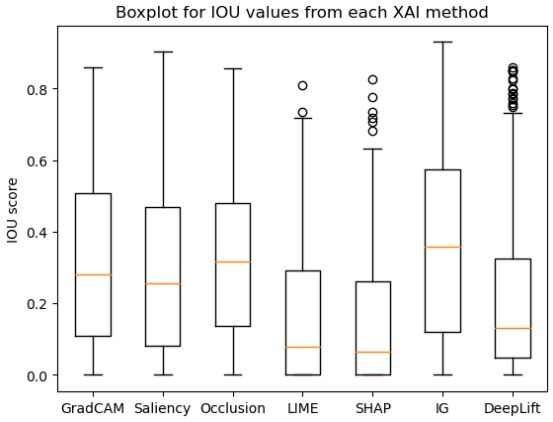


Figure 5: Boxplots computed with the IOU scores of the bounding boxes generated with the attribution maps of all the images. Higher IOU scores indicate a better overlap with the ground truth boxplots for lesion localization.

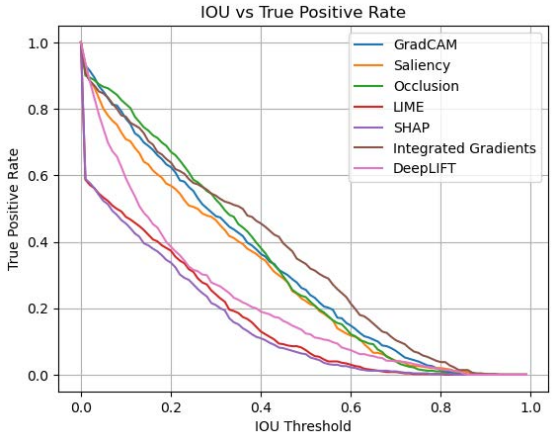


Figure 6: IOU vs TPR curves.

Fig. 3 illustrates each of the steps to generate the bounding boxes from the attribution maps

4. Results

4.1. Patch Classification

Attribution maps for several instances of the validation dataset were acquired in order to observe ResNet-50's performance and attempt to visualize its learned features. The randomly selected images indicate the prediction and the label, and each attribution map shows the attribution scores for each region of the image. The higher the score, the greater the impact of that region in the probability score for the target class. The attribution maps are color coded such that the redder a certain pixel, the higher its importance for the prediction, and the bluer, the lower its importance.

Fig. 4 shows the five original images plus their attribution map computed with each explainability method: Saliency, Integrated Gradients, Occlusion, Guided GradCAM, LIME, and SHAP. The maps differ greatly from one another, although some of them do seem to focus on areas that correspond to the lesion in the true positive cases. The true negative examples, however, visually vary more among themselves. On many cases, the focus of the model seems to be on regions which appear to be denser.

4.2. Full Image Classification

As with explainability for the patch classification task, the attribution maps for several images from the validation set were acquired. The true positive images with the highest probability scores were retrieved and their corresponding attribution maps for each of the methods were obtained. Figures 8 and 9 show the five images with the highest probability scores for malignancy, as well as the attribution maps generated with the explainability methods. All of the methods seem to highlight the areas corresponding to the lesions. Vanilla saliency appears to highlight areas outside the lesions more than the rest of the methods, but it still focuses mostly on the lesion.

The bounding boxes for each attribution map for each image per method were also generated. Fig. 10 and Fig. 11 show the corresponding bounding boxes generated from the attribution maps from Figs. 8 & 9, as well as the ground truth bounding boxes.

Examples showing the images with the lowest true positive scores and false negative scores were also generated, and are added in the appendix. In the case of the lowest true positive score examples, even though the network was not very confident at classifying the images, it does highlight the area corresponding to the lesion in most of the cases. The highest false negative score examples displayed this same behavior, focusing in the lesion area or highlighting it along with other regions in many cases.

Once having generated these examples for visual appraisal, the IOU for all the validation images for all the methods were computed as described in the previous section. Bounding boxes were computed to compare the overall IOU scores among the different explainability techniques, and are shown in Fig. 5. The best performing XAI techniques in terms of IOU scores were Integrated Gradients and Occlusion, while LIME and SHAP were found to be the lowest performing ones.

An IOU threshold vs True Positive Rate graph was also generated for further comparison of the seven different methods by varying the IOU thresholds. A TP would mean the bounding boxes from the attribution maps overlap significantly with the ground truth, and are therefore able to locate the lesions. The graph is shown in Fig. 6. As before, the best curves correspond to Integrated Gradients and Occlusion, while the worst belong to LIME and SHAP.

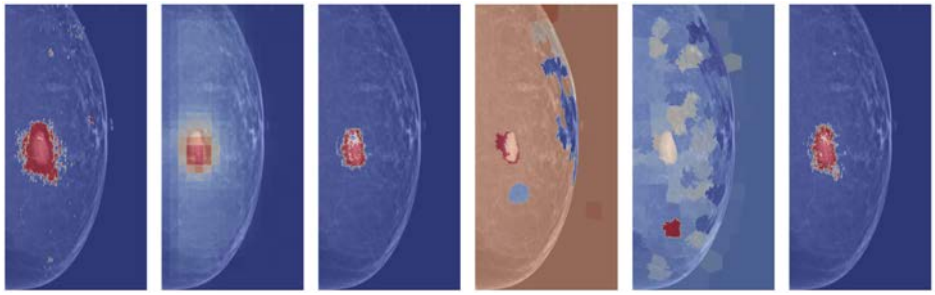


Figure 7: Attribution maps for the synthetic lesion generated over a real mammogram via stable diffusion.

4.3. Synthetic Lesions with Stable Diffusion

We have developed an additional experiment to further test attribution maps in synthetically generated lesions in normal mammograms. This experiment has been done in collaboration with Montoya (2023), whose master thesis work focused on the generation of high-resolution synthethic mammographies with diffusion models, and is featured in this year's proceedings. A sample image of a real mammogram with a synthetic lesion generated during his work was selected to extract attribution information with some explainability methods, as shown in Fig. 7. Most of the methods assign the synthetic lesion's area with the highest attribution scores. SHAP was the only method where this did not happen, although the synthetic lesion was counted as being of moderate impact.

5. Discussion

In this study, several explainability methods were tested on two networks, each trained on one of two breast cancer classification tasks. The explainability maps generated from the patch classifier exhibit an apparent overlap with the area of the patch corresponding to the mass, in the case of patches with masses. In general, the attribution maps seem to highlight the denser areas of the patches. The attribution maps look quite different from one another. This was to be expected for non-malignant patches, as, with the absence of a lesion, there is nothing that the classifier should focus on in particular. For the patches containing a lesion, it was observed that most of the attribution methods higher scores were often positioned in the middle of the patch. All in all, patches with no lesions seem to indicate that the network's attention seems to be dispersed throughout the image, whereas lesion patches' attribution maps mostly coincide with the lesion's position. This suggests that the network apparently learned to identify the masses properly, although it is difficult to ascertain when many lesions span almost the entirety or the majority of the patch, thus generating rather disperse attribution maps. Since the dataset consisted of patches centered around the lesion, this behavior is reasonable, as almost the entire image is visually different from normal patches and most of the the image could provide important information.

In the whole mammogram classification task, the attribution maps for all of the methods were successfully acquired for the validation dataset. When compared to the attribution maps from the previous task, the focus on specific regions of the image was much more apparent.

In true positive images, the highlighted regions mostly coincided with the position of the lesion in all the methods. Vanilla saliency presented more disperse attribution maps, presenting generally more clusters of highlighted pixels than the rest. This could be cause by vanilla saliency not being class discriminative, thus highlighting areas that contribute negatively towards the malignant target class. SHAP and LIME, for their part, although correctly assigning the highest attribution scores to the superpixels that overlapped with the lesion, occasionally allocated relatively high scores to superpixels that were positioned elsewhere in the image. A reason for this could be that those superpixels overlapped with several pixels that got high attribution scores, which, by themselves, would not be as noticeable, but by congregating in a superpixel the latter's attribution scores would rise and be more visually apparent across a larger area.

The bounding boxes generated from the attribution maps were overall able to center on the region indicated by the ground truth, although the large majority of them were quite bigger than their ground truth counterparts. This was expected, as they were acquired from the previously dilated attribution maps, which enlarged the highlighted regions. Integrated gradients achieved the highest scores, but generating their attribution maps was the computationally expensive and took the longest out of the seven methods. In contrast, even though Occlusion and GradCAM did not achieve IOU scores as high as Integrated Gradients, they were much faster, with GradCAM's attribution maps generating almost instantly. LIME and SHAP's much lower IOU scores were mostly caused by their bounding boxes being affected by the image's initial segmentation. Segmenting the mammogram into smaller regions could possibly help mitigate this, although it would increase the computationally resources when acquiring the attribu-

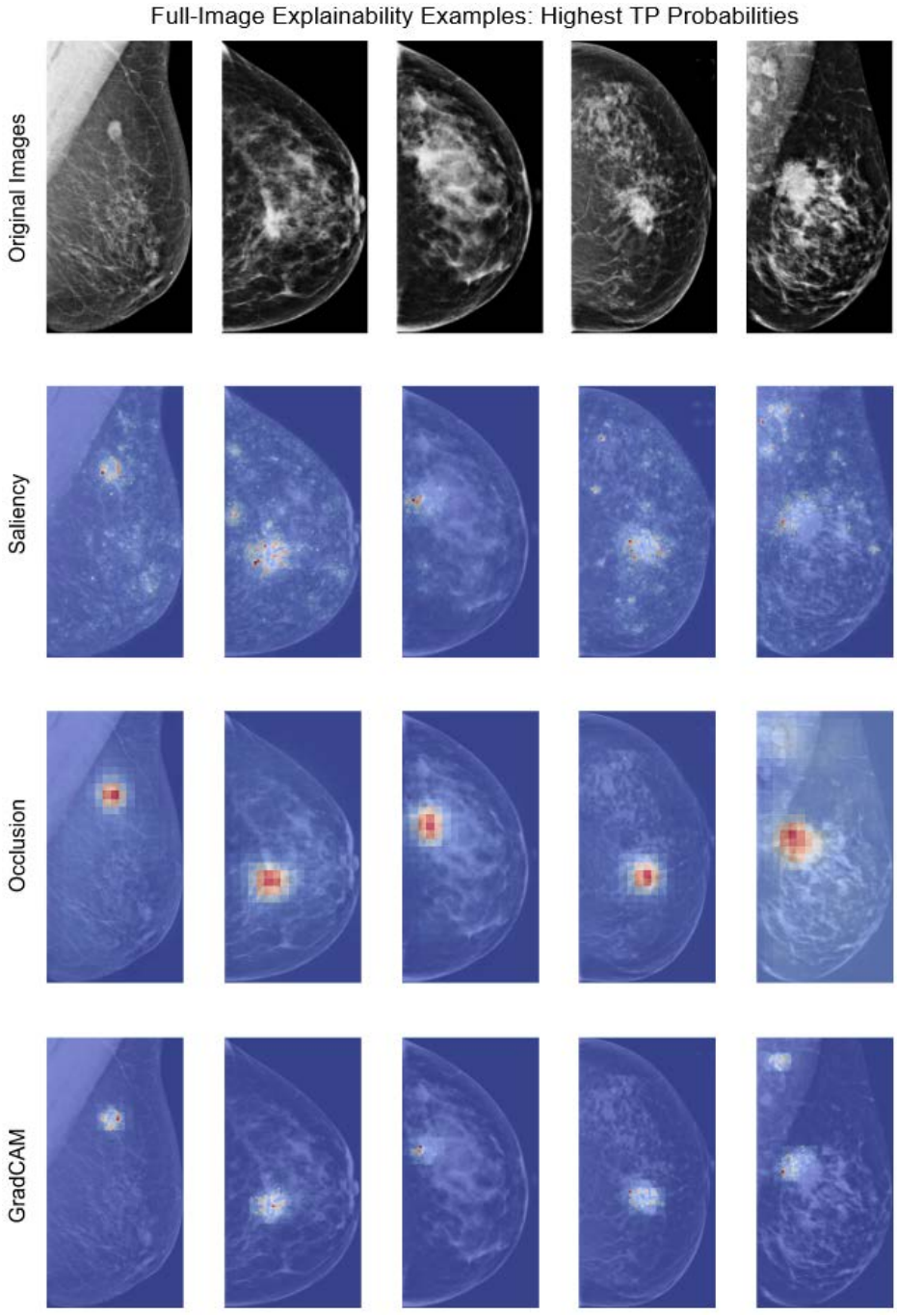


Figure 8: Examples for Saliency, Occlusion, and GradCAMs' attribution maps on images with high probability scores. Red and yellow regions correspond to higher attribution scores.

tion scores, as LIME and SHAP would have now many more image segments to consider.

True positive images with low probability scores and false negative images generated attribution maps where most of them contained the lesion area among its highlighted regions, indicating that while the model was not as confident classifying it as a malignant image or even misclassified it altogether, it was still focusing in the lesion. This supports the idea that the network successfully learned the masses' features and effectively bases its decision on the relevant data. The bounding boxes in this cases were expectedly not as accurate as the ones generated from the true positive images with higher probability scores, as even though the mass' position was highlighted, in many instances this was only one of several highlighted regions, which effectively created more bounding boxes throughout the mammogram, This resulted in a much bigger final bounding box when combining them and in predictably much lower IOU scores.

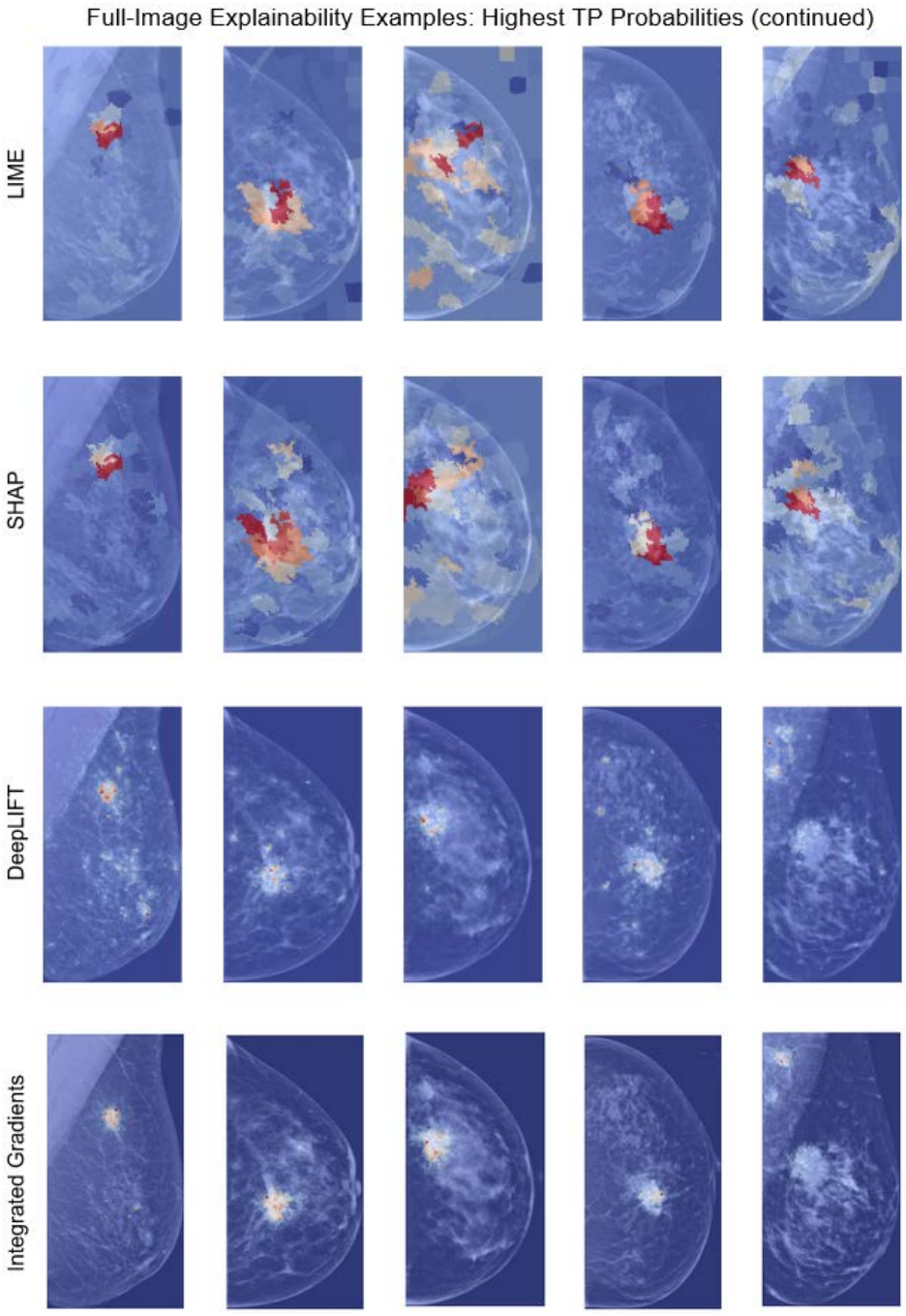


Figure 9: Examples for LIME, SHAP, DeepLIFT, and Integrated Gradients' attribution maps on images with high probability scores. Red and yellow regions correspond to higher attribution scores.

Finally, the attribution maps from the synthetic lesion generated with stable diffusion highlighted the lesion's area. This means that even while being synthetic, the explainability methods indicate that the network was focusing on it when evaluating for malignancy. Therefore, explainability algorithms could provide some insights for evaluating stable diffusion results.

5.1. Limitations and Future Work

The models trained for these classification tasks were not the best performing ones, which most certainly impacted the attribution maps. It could be possible to obtain attribution maps which highlight the lesions even more accurately if applied to better performing models. The bounding boxes were generated following a simple and rudimentary technique, causing them to be larger and less accurate. A more refined method to generate the bounding boxes could be developed, which could

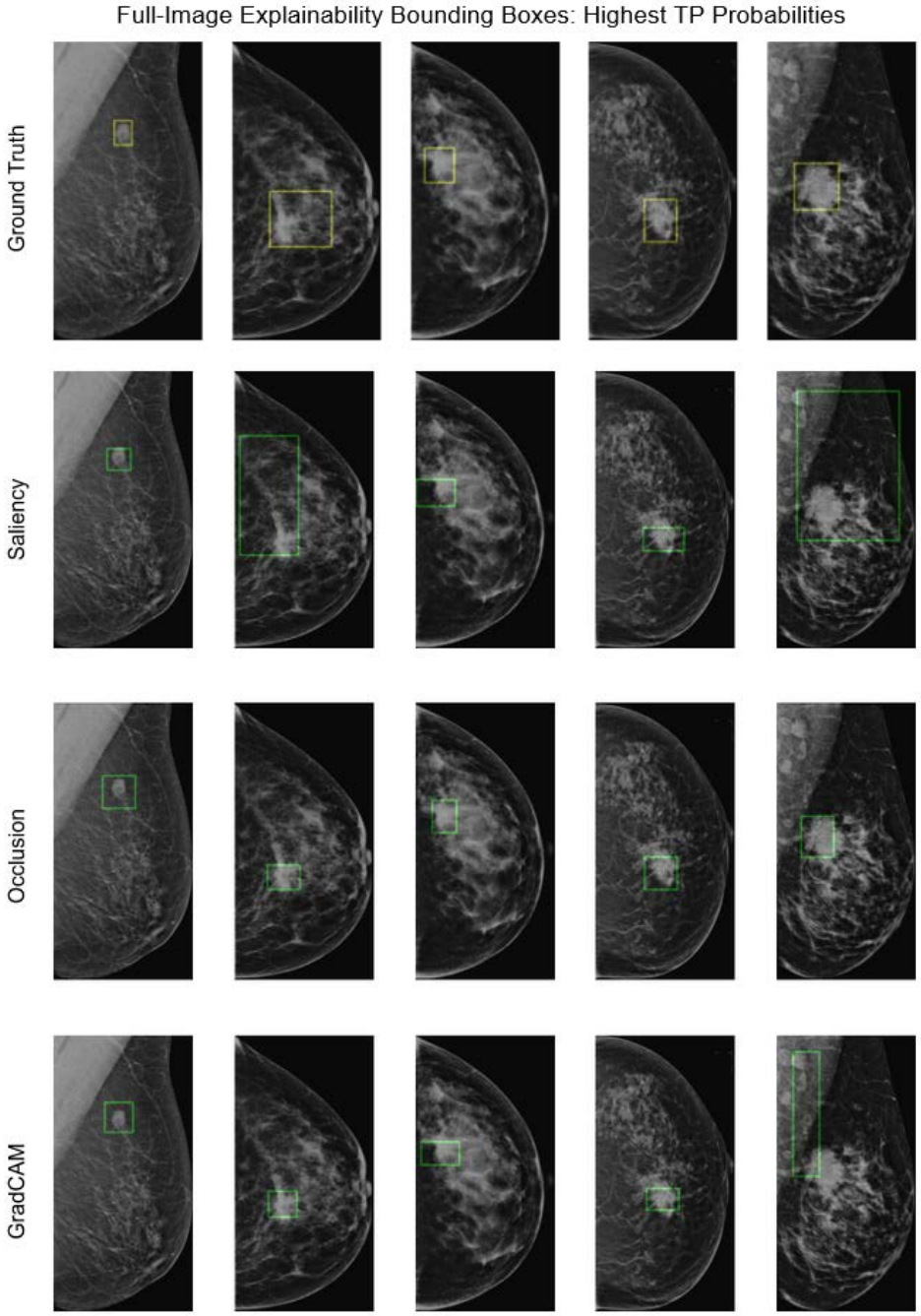


Figure 10: Examples of bounding boxes obtained with Saliency, Occlusion, and GradCAMs' attribution maps on images with high probability scores.

in turn improve the IOU results. Finally, a decently accurate model that classifies between different breast cancer subtypes could provide useful information in the attribution maps. At present, it is not possible to identify these subtypes without a more invasive procedure, but XAI could potentially provide useful morphological information for detecting these subtypes in mammograms.

6. Conclusions

In this paper, some XAI techniques were applied on two breast cancer classification tasks. The results obtained indicate that these techniques could provide users with useful information for understanding the decision-making processes of a neural network in medical imaging. When correctly trained, the methods should highlight the areas with clinical relevance, which in this case translated to the lesions in malignant mammograms. They can also show possible areas of improve-

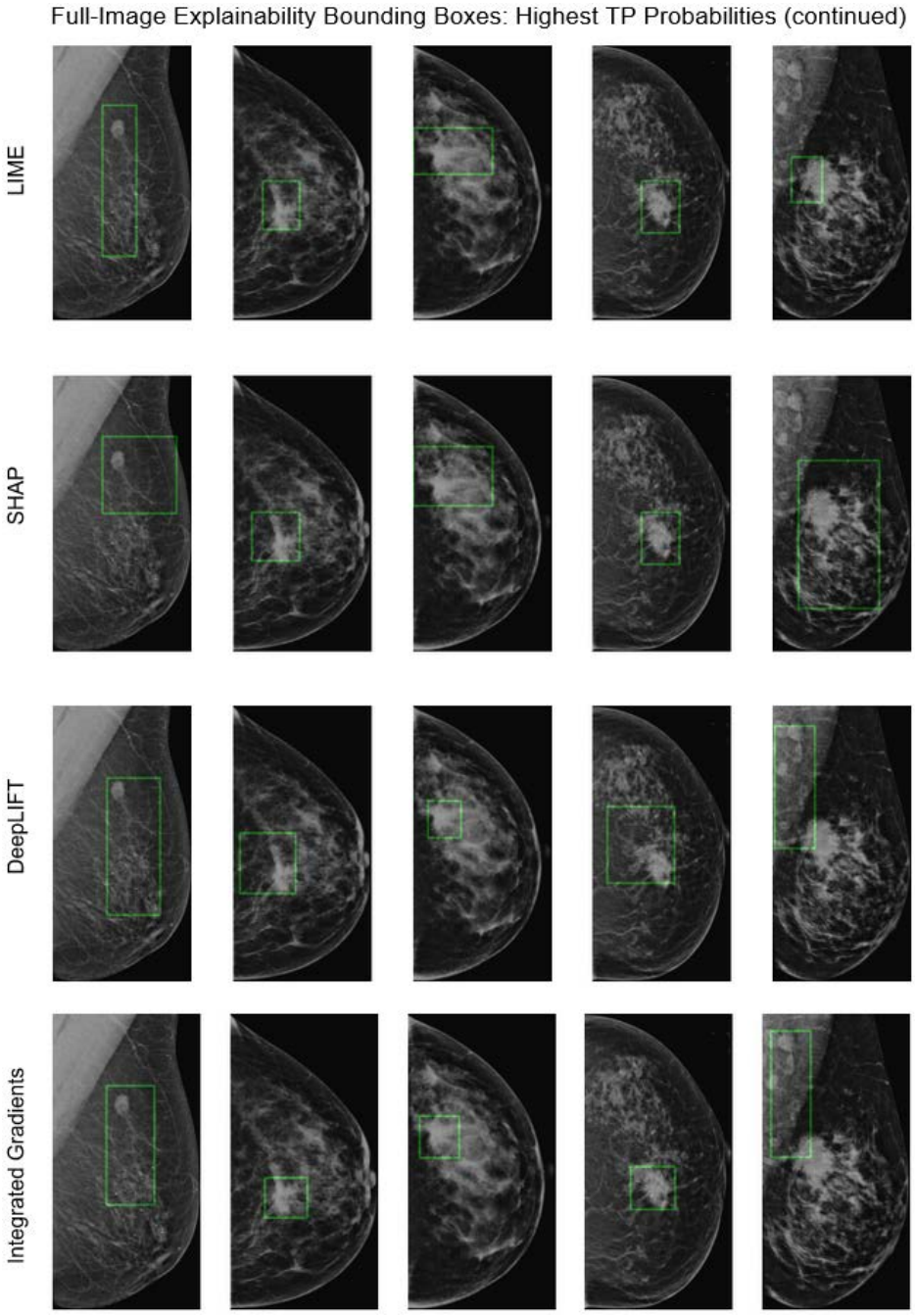


Figure 11: Examples of bounding boxes obtained with LIME, SHAP, DeepLIFT, and Integrated Gradients' attribution maps on images with high probability scores.

ment by indicating the regions which are causing the network to misclassify the images, or to detect possible biases. IOU scores with the lesion location were low but could potentially increase by improving the bounding box generation method from the attribution maps. Integrated gradients possessed the best IOU scores, but it is a rather computationally expensive technique. Grad-CAM and Occlusion could be used instead with potentially slightly worse results. Additionally, the various explainability techniques possess a potential for providing insights for evaluating and subsequently improving the models for the generation of synthetic images via stable diffusion. Future work could improve on this by refining the bounding box generation from the attribution maps and by applying XAI methods for breast cancer subtype classification.

Acknowledgments

I would like to express my heartfelt gratitude to my supervisor, Professor Robert Martí, for his invaluable

guidance and unwavering patience throughout the entire process of developing this thesis. Additionally, I extend my heartfelt appreciation to my fellow MAIA classmates for their continued assistance over the course of these two years, and for enriching this program with priceless experiences that will be remembered fondly. Lastly, I am deeply grateful to my family and friends, whose inconditional support made this journey possible.

References

Akserlod-Ballin, A., Chorev, M., Shoshan, Y., Spiro, A., Hazan, A., Melamed, R., Barkan, E., Herzel, E., Naor, S., Karavani, E., Koren, G., Goldschmidt, Y., Shalev, V., Rosen-Zvi, M., Guindy, M., 2019. Predicting breast cancer by applying deep learning to linked health records and mammograms. Radiology 292, 182622. doi:10.1148/radiol.2019182622.

Al-antari, M.A., Han, S.M., Kim, T.S., 2020. Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital x-ray mammograms. Computer Methods and Programs in Biomedicine 196, 105584. doi:https://doi.org/10.1016/j.cmpb.2020.105584.

Arai, S., Nagao, T., 2017. Intuitive visualization method for image classification using convolutional neural networks. Information Processing Society of Japan. Transactions on mathematical modeling and its applications 10, 1–13.

Arevalo, J., González, F.A., Ramos-Pollán, R., Oliveira, J.L., Guevara Lopez, M.A., 2015. Convolutional neural networks for mammography mass lesion classification, in: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 797–800. doi:10.1109/EMBC.2015.7318482.

Balkenende, L., Teuwen, J., Mann, R.M., 2022. Application of deep learning in breast cancer imaging. Seminars in Nuclear Medicine 52, 584–596. doi:https://doi.org/10.1053/j.semnuclmed.2022.02.003. breast Cancer.

Cantone, M., Marrocco, C., Tortorella, F., Bria, A., 2023. Convolutional networks and transformers for mammography classification: An experimental study. Sensors (Basel) 23.

Carr, C., Kitamura, F., Partridge, G., inversion, Kalpathy-Cramer, J., Mongan, J., Lavender, K.A., Vazirabad, M., Riopel, M., Ball, R., Dane, S., Chen, Y., 2022. Rsna screening mammography breast cancer detection. URL: https://kaggle.com/competitions/rsna-breast-cancer-detection.

Halling-Brown, M.D., Warren, L.M., Ward, D., Lewis, E., Mackenzie, A., Wallis, M.G., Wilkinson, L., Given-Wilson, R.M., McAvinchey, R., Young, K.C., 2020. Optimam mammography image database: a large scale resource of mammography images and clinical data.

Huang, Z., Zhu, X., Ding, M., Zhang, X., 2020. Medical image classification using a light-weighted hybrid neural network based on pcanet and densenet. IEEE Access 8, 24697–24712. doi:10.1109/ACCESS.2020.2971225.

Kobayashi, T., Haraguchi, T., Nagao, T., 2022. Classifying presence or absence of calcifications on mammography using generative contribution mapping. Radiological Physics and Technology 15. doi:10.1007/s12194-022-00673-3.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., Reblitz-Richardson, O., 2020. Captum: A unified and generic model interpretability library for pytorch.

Loizidou, K., Elia, R., Pitris, C., 2023. Computer-aided breast cancer detection and classification in mammography: A comprehensive review. Computers in Biology and Medicine 153, 106554. doi:https://doi.org/10.1016/j.compbiomed.2023.106554.

Ou, W.C., Polat, D., Dogan, B.E., 2021. Deep learning in breast radiology: current progress and future directions. European radiology 31, 4872—4885. doi:10.1007/s00330-020-07640-9.

Prodan, M., Paraschiv, E., Stanciu, A., 2023. Applying deep learning methods for mammography analysis and breast cancer detection. Applied Sciences 13, 4272. URL: http://dx.doi.org/10.3390/app13074272, doi:10.3390/app13074272.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "why should i trust you?": Explaining the predictions of any classifier.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2019. Grad-CAM: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision 128, 336–359. doi:10.1007/s11263-019-01228-7.

Shrikumar, A., Greenside, P., Kundaje, A., 2019. Learning important features through propagating activation differences.

Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps.

Singh, A., Sengupta, S., Lakshminarayanan, V., 2020. Explainable deep learning models in medical image analysis. Journal of Imaging 6.

Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic attribution for deep networks.

van der Velden, B.H., Kuijf, H.J., Gilhuijs, K.G., Viergever, M.A., 2022. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. Medical Image Analysis 79, 102470. doi:https://doi.org/10.1016/j.media.2022.102470.

de Vries, B.M., Zwezerijnen, G.J.C., Burchell, G.L., van Velden, F.H.P., Menke-van der Houven van Oordt, C.W., Boellaard, R., 2023. Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review. Front Med (Lausanne) 10, 1180773.

Xi, P., Guan, H., Shu, C., Borgeat, L., Goubran, R., 2020. An integrated approach for medical abnormality detection using deep patch convolutional neural networks. The Visual Computer 36. doi:10.1007/s00371-019-01775-7.

Yi, P., Lin, A., Wei, J., Yu, A., Sair, H., Hui, F., Hager, G., Harvey, S., 2019. Deep-learning-based semantic labeling for 2d mammography and comparison of complexity for machine learning tasks. Journal of Digital Imaging 32. doi:10.1007/s10278-019-00244-w.

Zeiler, M.D., Fergus, R., 2013. Visualizing and understanding convolutional networks. arXiv:1311.2901.

Appendix A. Attribution maps for whole mammograms

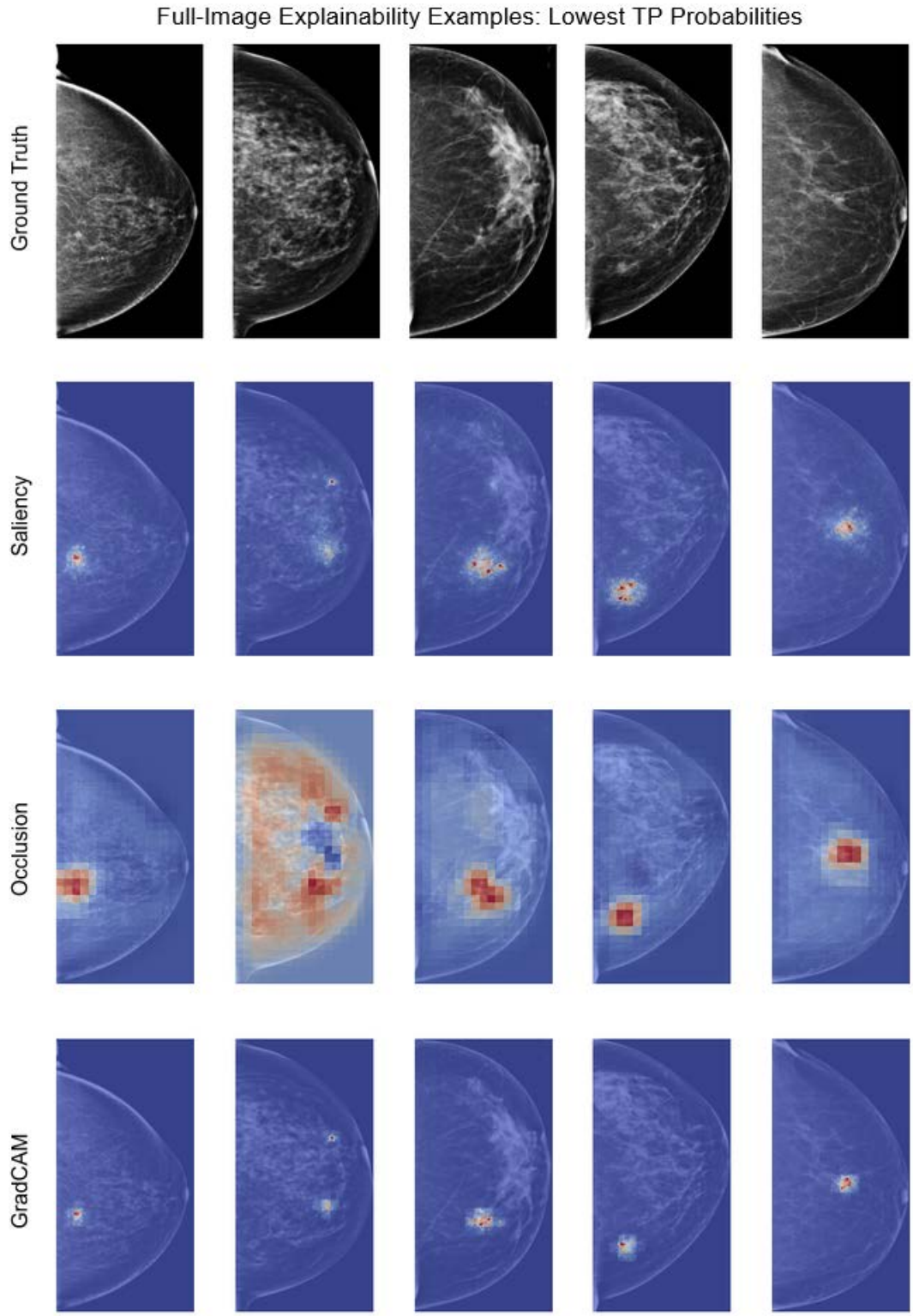


Figure A.12: Examples for Saliency, Occlusion, and GradCAMs' attribution maps on TP images with low probability scores. Red and yellow regions correspond to higher attribution scores.

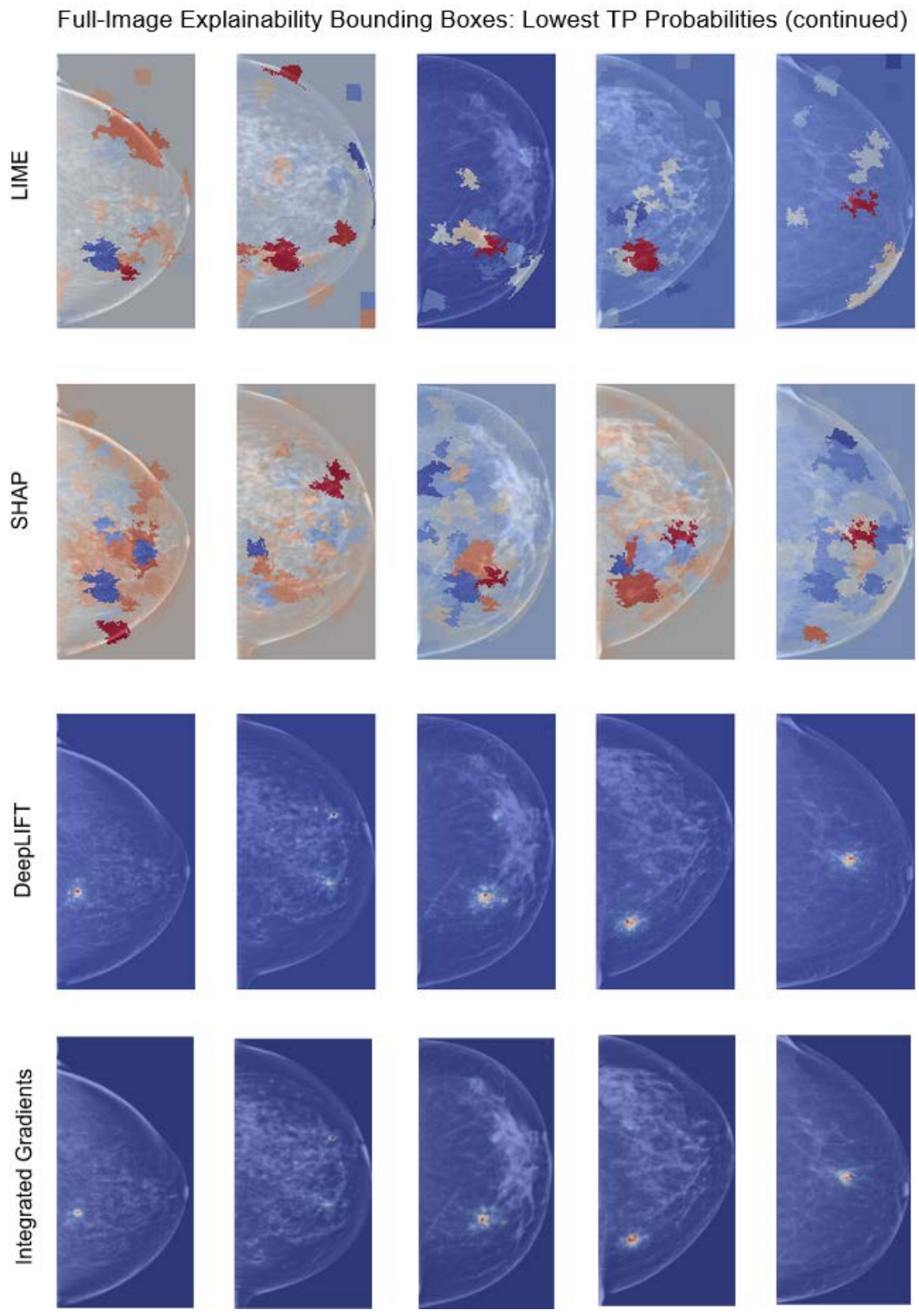


Figure A.13: Examples for LIME, SHAP, DeepLIFT, and Integrated Gradients' attribution maps on TP images with low probability scores. Red and yellow regions correspond to higher attribution scores.

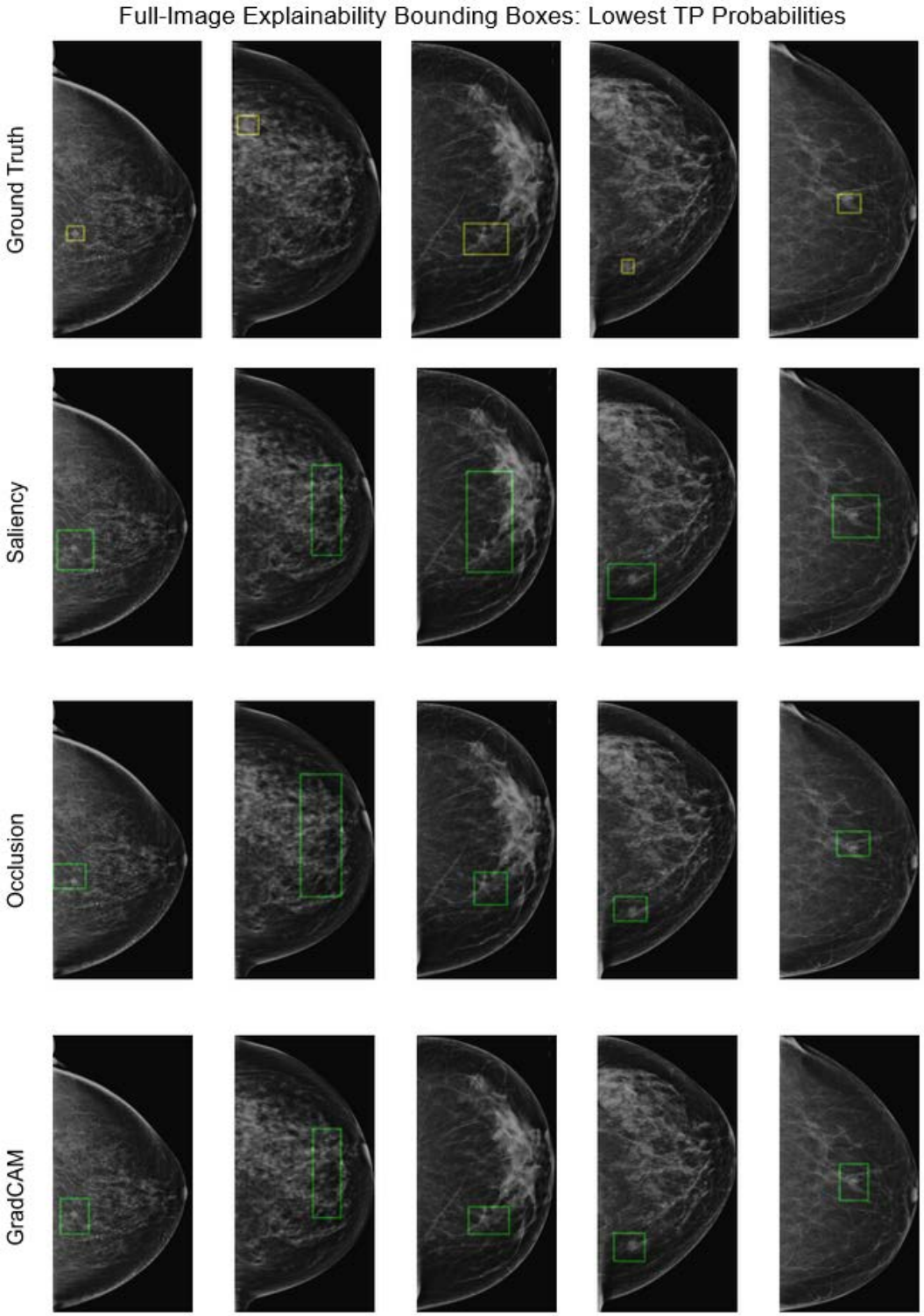


Figure A.14: Examples of bounding boxes obtained with Saliency, Occlusion, and GradCAMs' attribution maps on TP images with low probability scores.

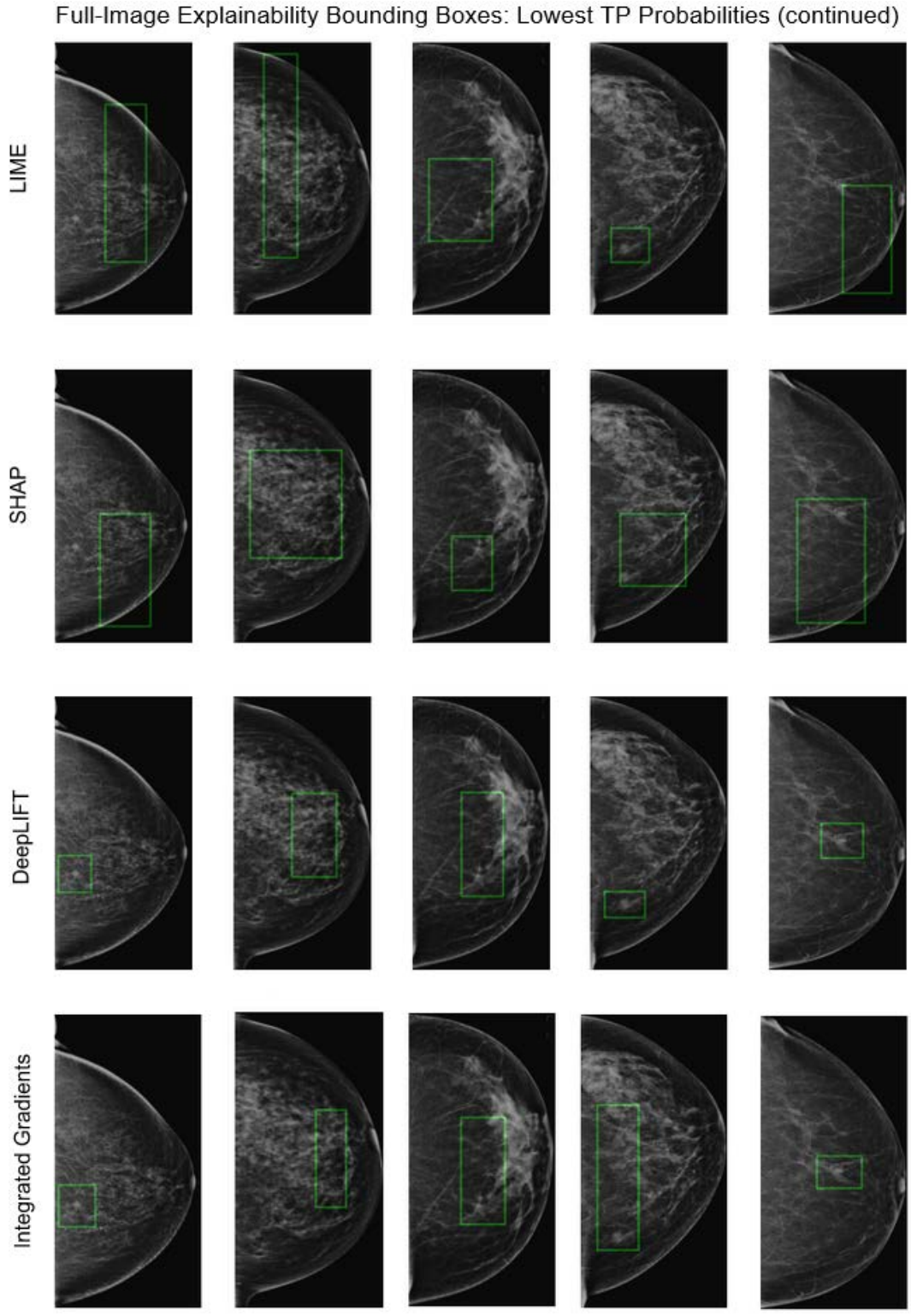


Figure A.15: Examples of bounding boxes obtained with LIME, SHAP, DeepLIFT, and Integrated Gradients' attribution maps on TP images with low probability scores.

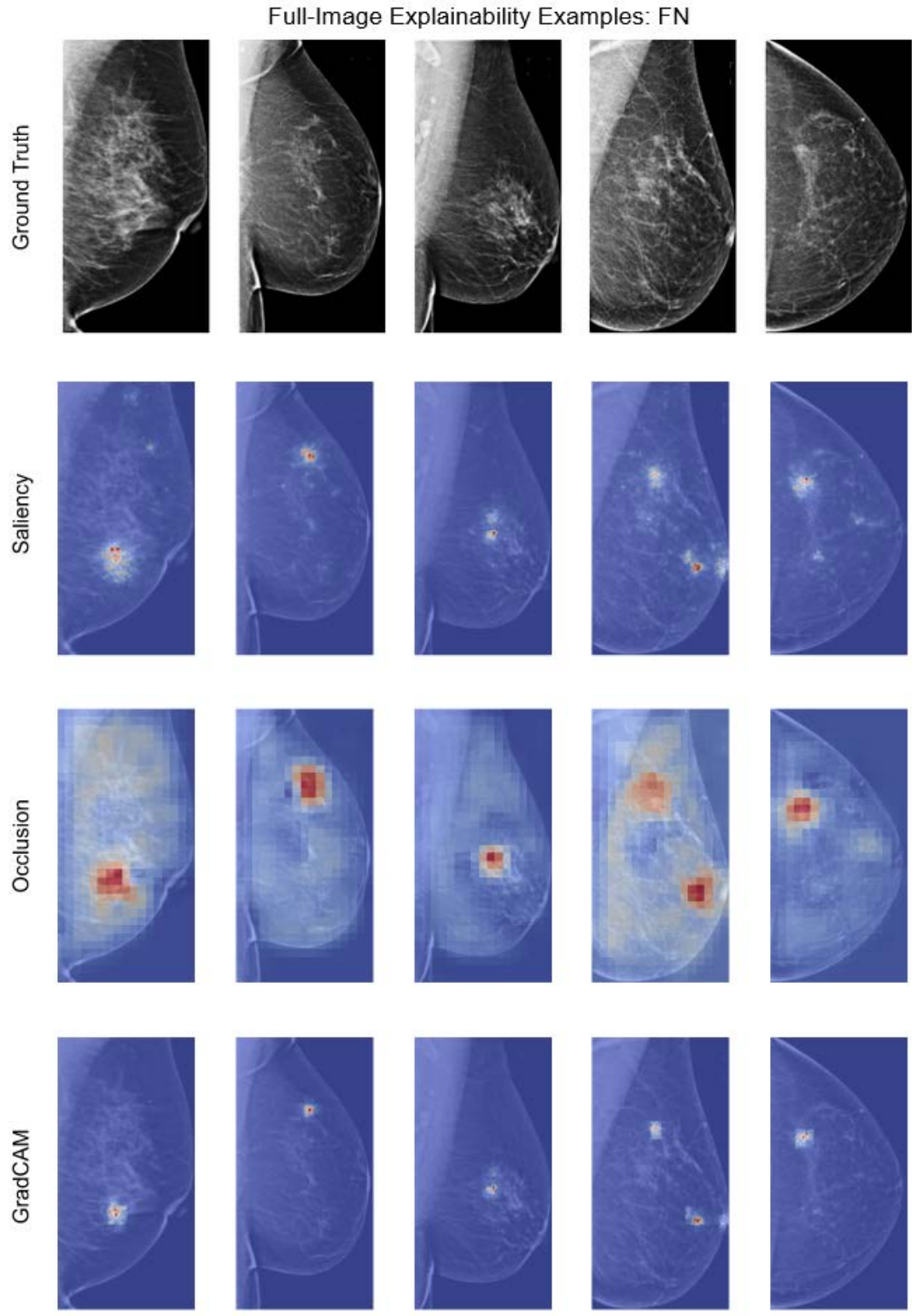


Figure A.16: Examples for Saliency, Occlusion, and GradCAMs' attribution maps on FN images. Red and yellow regions correspond to higher attribution scores.

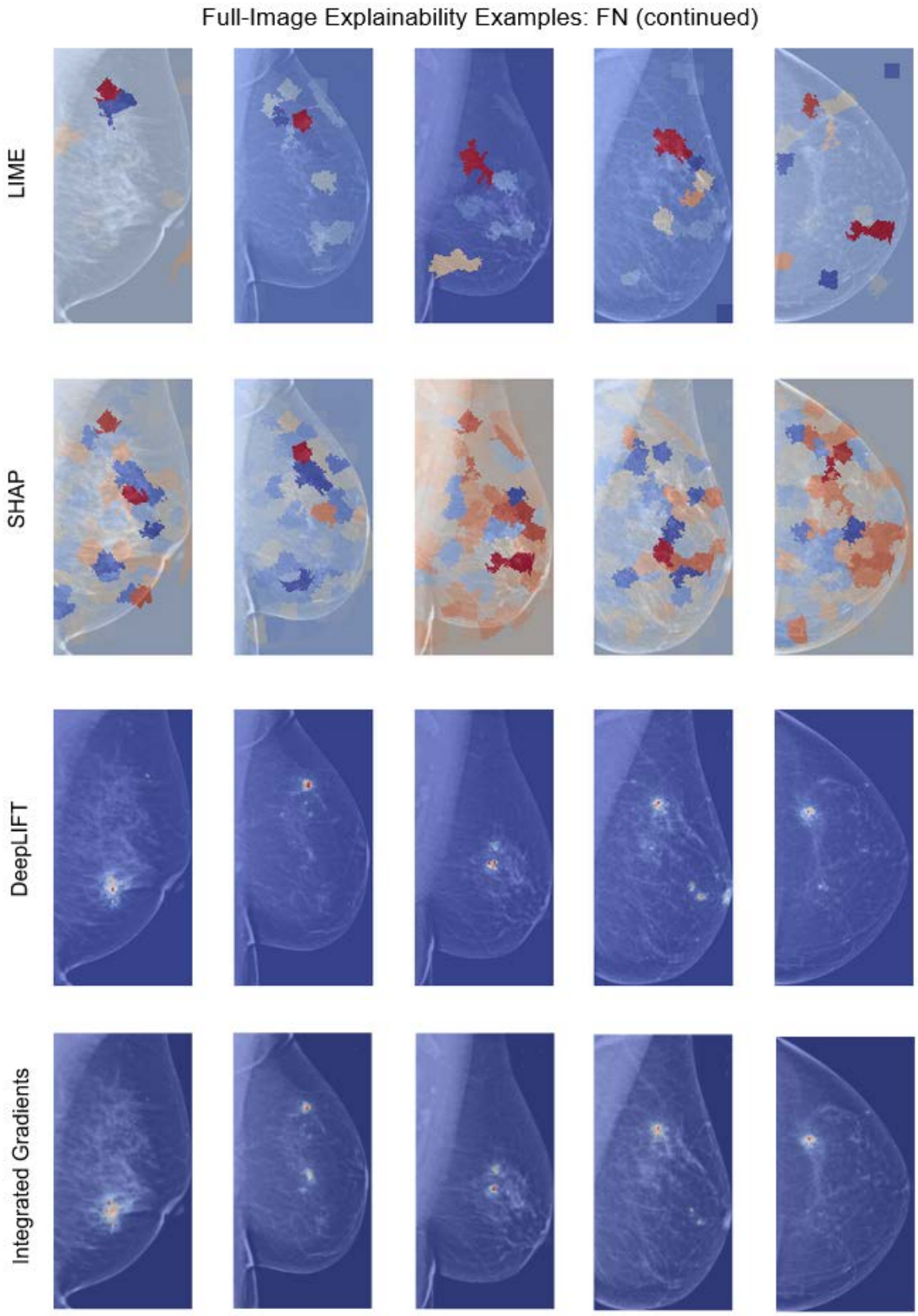


Figure A.17: Examples for LIME, SHAP, DeepLIFT, and Integrated Gradients' attribution maps on FN images. Red and yellow regions correspond to higher attribution scores.

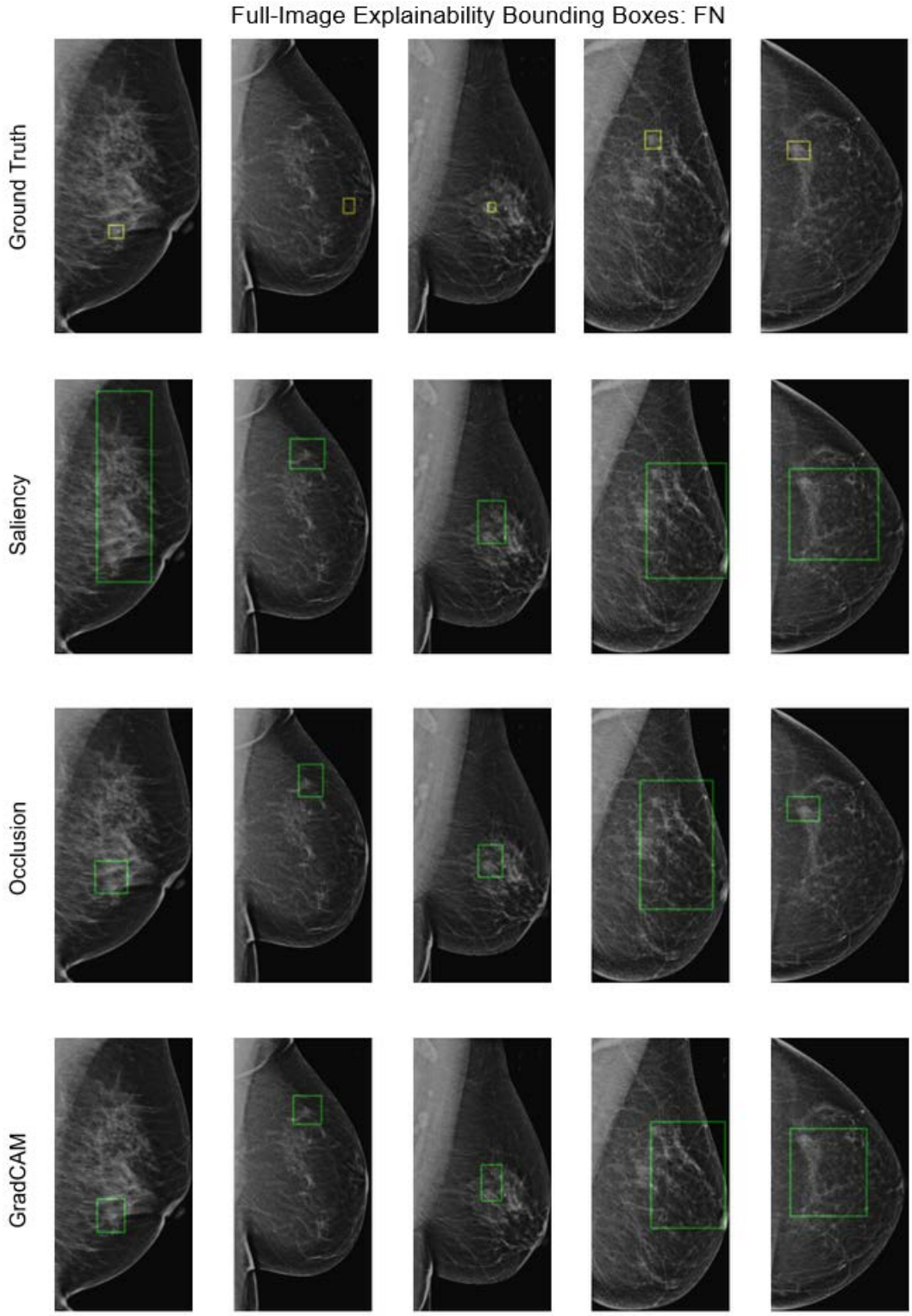


Figure A.18: Examples of bounding boxes obtained with Saliency, Occlusion, and GradCAMs' attribution maps on FN images.

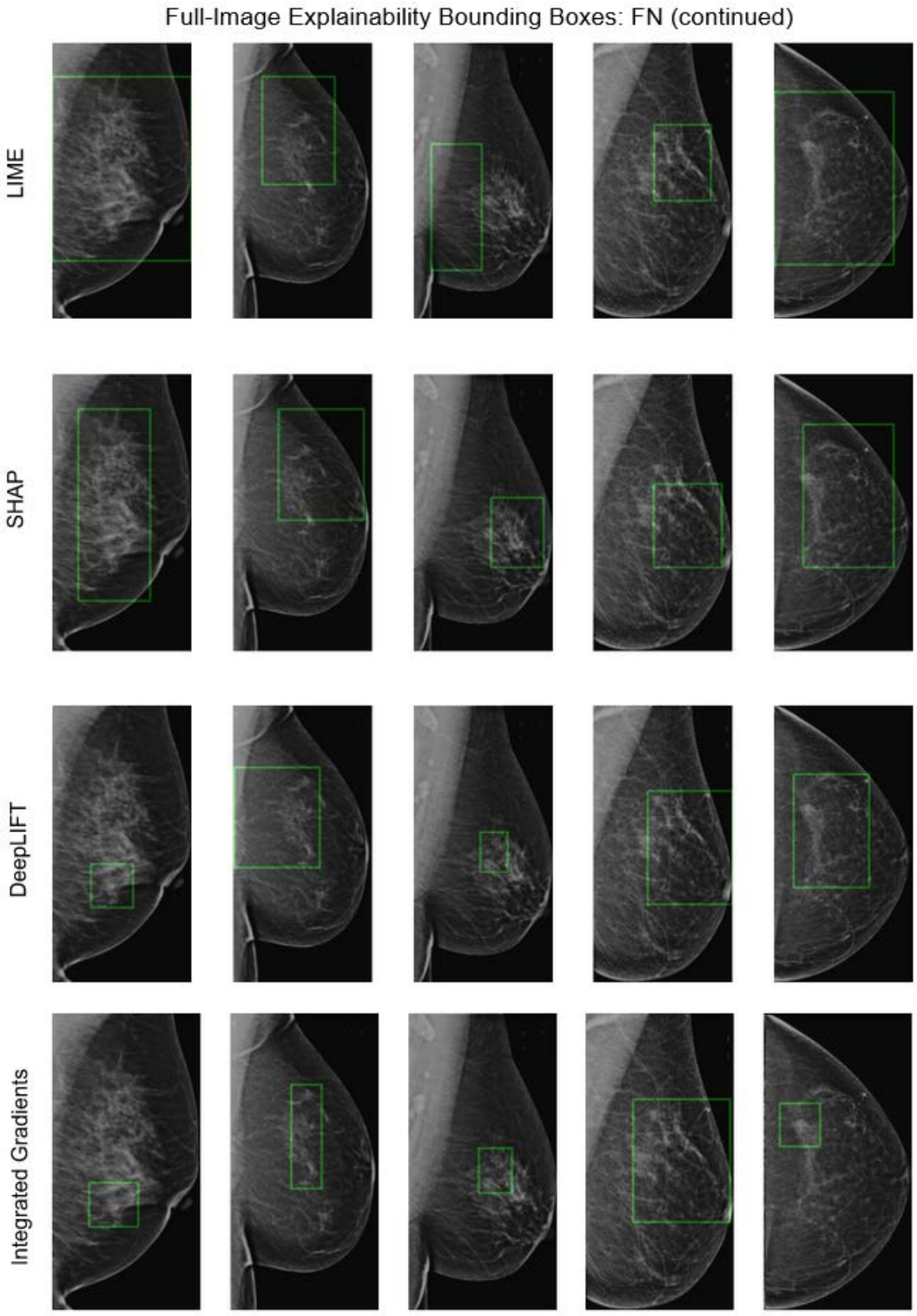
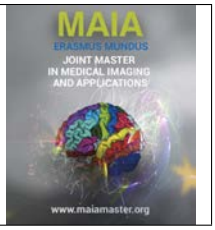


Figure A.19: Examples of bounding boxes obtained with LIME, SHAP, DeepLIFT, and Integrated Gradients' attribution maps on FN images.

Medical Imaging and Applications

Master Thesis, June 2023



Self-supervised learning for acute ischemic stroke final infarct lesion segmentation in non-contrast CT

Joaquin O. Seia*, Ezequiel de la Rosa[a], Diana M. Sima[a], David Robben[a]

[a] *icometrix, Leuven, Belgium*

Abstract

Ischemic stroke accounts for 87% of all strokes, which are the leading cause of disability and the fifth leading cause of death worldwide. Current stroke management guidelines rely on quantification of ischemic lesion volume to select an appropriate treatment for a patient. Despite the fact that baseline non-contrast CT is not as suitable as it is Perfusion CT or Diffusion Weighted Imaging to obtain this measurement, it is the first imaging modality performed when the patient arrives at the emergency department, it is cheaper, more widely available and faster. Consequently, the development of accurate automated segmentation tools for ischemic lesions on baseline non-contrast CT is a clinically relevant problem that, if satisfactory solved, would represent an improvement in healthcare provision.

Among other difficulties, the low contrast that the ischemic stroke lesion presents on baseline non-contrast CT makes the task of segmenting it very challenging even for expert radiologists. In the case of automated solutions, the difficulty of collecting large, well-curated datasets of baseline and follow-up acute ischemic stroke images further complicates the task. In this work, a self-supervised learning (SSL) pre-training strategy was proposed to exploit large unlabelled non-contrast CT datasets (stroke positive and negative) in the task of acute ischemic stroke infarct segmentation. A robust data pre-processing pipeline was proposed to homogenise the different datasets before using them in a SSL-enhanced version of the well-known self-configuring nnU-Net pipeline. From the experiments conducted, pre-training the nnU-Net encoder in a self-supervised manner with all available non-contrast CT images resulted in an acute ischemic stroke segmentation performance significantly higher than training the same model from scratch and comparable to that obtained by training from scratch using approximately 3.6 times more labelled data.

The code developed for this work is publicly available at: https://github.com/joaco18/stroke-seg-ssl.

Keywords: Acute Ischemic Stroke, Non-Contrast CT, Segmentation, Self-Supervised Learning

1. Introduction

Stroke is a pathology characterised by a focal injury in the central nervous system with a vascular origin. It represents the first cause of disability and the fifth cause of death worldwide (Virani et al. (2021)). A stroke can be classified in two types: ischemic and hemorrhagic. The ischemic type accounts for 87% of all strokes and involves a restriction or reduction of blood flow caused by the occlusion of a blood vessel (Benjamin et al. (2017); Sacco et al. (2013)).

During the management of this emergency, *time is brain*: the longer the brain tissue is deprived of blood supply, the higher the probability of cell death and the worse the prognosis for the patient. Depending on the time elapsed since the onset of the stroke, it can be divided into three categories: *hyperacute* (less than 6 hours from onset), *acute* (less than 24 hours) or *subacute* (from 24 hours to 5 days) (Brorson and Cifu (2019)). Each of these stages is accompanied by distinct physiopathological characteristics that define different ways in which the patient should be managed.

The time-dependent fate of the hypoperfused tissue can be spatially described by two clinically relevant zones: *core* and *penumbra*. While the former refers to a highly hypoperfused tissue that is already infarcted (or is inevitably destined to become infarcted regardless of treatment), the latter represents hypoperfused tissue

*Corresponding author
Email address: joacoseia18@gmail.com (Joaquin O. Seia)

that is potentially salvageable through reperfusion (Vagal et al. (2019)).

Reperfusion therapy, and endovascular treatment (EVT) in particular, is an effective therapeutic solution for acute ischemic stroke (AIS) patients with a large vessel occlusion. However, it is currently restricted to patients with a small lesion core because the larger this region it is, the higher the risk of an hemorrhage following the reperfusion and the smaller the benefit the patient can get from it. (Byrne et al. (2019); Goyal et al. (2020)).

1.1. Medical images in the context of stroke

As mentioned above, selecting the correct treatment for a patient requires the assessment of the extent of the ischemic lesion, and medical imaging plays a vital role in this process. This is reflected in the current American Heart Association (AHA) guideline for the management of stroke patients, which recommends an emergency brain imaging evaluation before any treatment decision is made (Powers et al. (2019)). This recommendation states that non-contrast computed tomography (NCCT) should be a first-line imaging modality used to rule out intracerebral hemorrhage. Thereafter, patient selection for EVT should follow one of two imaging recommendations depending on the time to last known well. For hyperacute ischemic stroke, the Alberta Stroke Program Early CT Score (ASPECTS) should be computed over the NCCT and an angiographic CT (CTA) should be acquired. In AIS patients, a combination of CTA and perfusion CT (CTP) or magnetic resonance angiography (MRA) and diffusion-weighted magnetic resonance imaging (DWI) is recommended.

The guideline presents three principal non-angiographic imaging modalities: NCCT, CTP and DWI. Although the evaluation of an AIS stroke patient using NCCT alone is not recommended, it is an imaging modality with high potential for stroke lesion assessment. In order to identify what makes NCCT unique, it is necessary to introduce on some basic concepts of these three modalities.

Non-contrast computed tomography

NCCT measures tissue density. Brain tissue undergoing through severe ischemia appears hypodense on NCCT because of increased water content due to ionic edema (Goyal et al. (2020)). ASPECTS is a rating system that uses this biomarker to subjectively assess the extent of early infarction on the NCCT (Mokin et al. (2017)). Although the use of ASPECTS is currently part of the stroke management guidelines, it is characterised by a high inter-observer variability (Farzin et al. (2016)). It has been well described that the ischemic core signal is virtually absent in baseline NCCT images compared to other modalities, making the task of lesion segmentation challenging even for expert neuroradiologists (El-Hariri et al. (2022); Estrada et al. (2022)). Furthermore, this scoring system does not provide a fine quantification of the lesion volume but rather a coarse estimation of extent based on affected vascular regions.

Computed tomography perfusion

In this modality, the focus is placed on blood flow measurement rather than the consequences of ischemia in the brain parenchyma. Through a non-trivial post-processing step, measurements of penumbra and ischemic core volumes can be obtained based on the relative blood flow at each voxel. The recent inclusion of CTP among the AHA recommendations results from the successful use of CTP-derived core and penumbra volumes as part of patient selection criteria for EVT in two large clinical trials, Defuse 3 and DAWN (Albers et al. (2018); Nogueira et al. (2018)).

However, despite being useful, CTP is not free of complications. Differences in results between software solutions and difficulties inherent in the modality itself, such as patient motion or confounding physiological processes, can lead to over- or underestimation of core volume in CTP. To serve as an example, in AIS, CTP is prone to underestimation of baseline ischemic core in cases of *luxury perfusion*, where the core infarction becomes hyperemic because of spontaneous reperfusion or engorgement of the leptomeningeal arteries (Sotoudeh et al. (2019)). In order to rule out this false negatives, a simultaneous review of the baseline NCCT is required, leveraging the complementary information provided by the two modalities (Vagal et al. (2019)).

Diffusion weighted image

The ischemic stroke lesion appears as a high signal on the DWI scan because of the diffusion restriction in the extracellular space caused by the cytotoxic edema (Goyal et al. (2020), Kuang et al. (2021)). Contrary to NCCT, this biomarker is visible within minutes after ischemia onset and is much more conspicuous. Because of its limited availability, the higher cost and longer acquisition time, DWI is usually reserved for a follow-up evaluation and quantification of final infarct (El-Hariri et al. (2022)). These elements make DWI the gold standard for estimating the volume of the ischemic lesion.

However, it is important to note that even though the lesion core volumes computed on the baseline NCCT (or CTP) are highly correlated with those obtained from the DWI image, they may differ. For example, depending on the success of recanalisation or the time elapsed between the baseline image and the treatment, the final infarct extent will differ from the baseline lesion. In addition, very small ischemic lesions associated with small emboli generated during reperfusion may appear on the post-treatment image but not on the baseline image.

1.2. Why NCCT?

As presented, it is clear that advanced imaging techniques such as CTP and DWI can provide a more com-

plete and accurate assessment of the ischemic lesion. However, these modalities are not available in hospitals on a 24/7 basis, and AIS patients are mostly diagnosed by using NCCT images (Kim et al. (2021)).

Despite the fact that baseline NCCT is not the best modality for quantifying the volume of the ischemic core, it is the first line imaging modality performed when the patient arrives to the emergency department, is the cheapest, the most widely available and the fastest technique among the mentioned ones. Consequently, the segmentation of ischemic stroke lesion core on baseline NCCT is a clinically relevant problem.

As stated in Bouslama et al. (2021), a proper quantification of the stroke core on baseline NCCT images could allow centres without advanced imaging techniques or specialised stroke neurologists to ensure access to endovascular therapy for a wider population of patients who could benefit from it. Even when following the current guidelines and using perfusion imaging, NCCT can still provide complementary information that can lead to improved healthcare provision.

In this context, where manual segmentation of stroke lesions on baseline NCCT images is not feasible but would represent an improvement in the management of AIS patients, the development of accurate automated segmentation tools for ischemic lesions on baseline NCCT is a problem that needs to be addressed.

2. State of the art

2.1. Automatic segmentation of AIS lesions

Over the last decade, machine learning, and in particular deep learning (DL), has been successfully applied to many image segmentation taks. The work in Isensee et al. (2020), which presented a robust model that achieved state of the art (SOTA) performance over 53 different medical image segmentation problems, can serve as a clear example of this. The segmentation of acute ischemic lesions has not been the exception, where several methods have tackled the task in MRI images (Clèrigues et al. (2020)) or CTP scans (Amador et al. (2021, 2022); Robben et al. (2020)).

In contrast, on baseline NCCT images, there are only a few well-established approaches for segmenting AIS lesions. Among these solutions, there is a large heterogeneity in their experimental designs, which affects how comparable and transferable they are to other NCCT datasets. Overall, there are three main elements transversal to the literature on this topic:

1. The vast majority of deep learning proposals have successfully applied UNet-like deep convolutional neural networks (DCNN).
2. The use of contextual information in the model design improves the segmentation results. In most cases, inter-hemispheric asymmetries are used as one of the forms of contextual information.
3. The lack of large, well-curated and publicly available datasets containing baseline and follow-up AIS images is not negligible, as most of the publications have worked with private datasets with different patient selection criteria.

In the following for each of this three aspects some salient publications are commented.

2.1.1. U-Net like architecture choice

Among the many works using U-shaped architectures, in Ostmeier et al. (2022) and El-Hariri et al. (2022), the self-configuring model nnU-Net (Isensee et al. (2020)) was shown to be successful in AIS lesion core segmentation on baseline NCCT images. In the first case, the authors showed that nnU-Net achieved non-inferior segmentation results compared to expert neuroradiologists. In the second case, the authors not only showed that nnU-Net was able to achieve high volumetric agreement with ground truth pre-treatment DWI labels, but also pointed out that their model was already part of commercial software, demonstrating the impact this architecture already has in the clinical practice.

2.1.2. Exploiting contextual information

Contextual information has been incorporated into models in many different ways in the literature. Chen et al. (2022) and Kuang et al. (2019) used the difference images generated after a sagittal flipping of the NCCT images. The first one opted for a 2D U-shaped architecture in which the original, flipped and difference images were given as a multi-channel input. The second one opted for a more sophisticated approach using a 3D U-shaped architecture with a multi-path encoder. Four paths were used, covering the original image, the difference image, an infarct location probability map and a distance-to-cerebrospinal-fluid map. In Ni et al. (2022), a 3-step end-to-end trainable 3D asymmetry disentangling network was used to obtain an effective and interpretable AIS segmentation on NCCT. Their method automatically separated pathological asymmetries and intrinsic anatomical asymmetries from the NCCT.

An approach usually referred to as SOTA in AIS core segmentation in baseline NCCT is the work of Kuang et al. (2021). The authors proposed a multi-task learning approach, called EIS-Net, which was simultaneously trained to segment the stroke lesion and to predict the ASPECTS score from the NCCT. Their model consisted of a 3D U-shaped segmentation CNN architecture with a triple-path encoder. Each path was fed with the NCCT, the sagittally mirrored NCCT and a CT atlas, respectively. The differences between these features were exploited using an ad hoc comparison block. The use of contextual information within a multitask optimisation strategy allowed them to achieve better results than using the plain U-shaped segmentation model.

2.1.3. Data scarcity problem

Overlapping with the previous remark, the work in Giancardo et al. (2023) presents another model that exploits the inter-hemispheric differences. However, the remarkable aspect of this paper is that the authors identified that one of the elements that is delaying the development of automatic AIS segmentation in NCCT/CTA is the difficulty in obtaining large enough samples containing high-quality DWI images with voxel-level ground truth annotations. To circumvent this problem, they used only image-level labels (the stroke core volume size) to train their model and obtained a competitive AIS lesion core segmentation on CTA images.

2.2. Deep learning with labelled data scarcity

In the recent years, self-supervised learning (SSL) strategies have gained popularity for addressing the problem of scarcity of labelled data. As described in the work of Balestriero et al. (2023) SSL stands for a collection of machine learning approaches that can learn from large amounts of unlabelled data. The common practice implies the definition of a pretext task based on unlabelled inputs to produce descriptive and meaningful representations that can be used across different downstream tasks. Self-supervised image representation learning has shown amazing progress in the last five years, achieving a performance in several downstream tasks that is competitive or even superior to supervised learning approaches (Bardes et al. (2022); Caron et al. (2021); Chen and He (2021); Grill et al. (2020); He et al. (2022, 2020)).

One of the most commonly used SSL methods is based on a joint embedding architecture, where two Siamese networks are trained to produce similar embeddings for different views of the same image. In this way, the networks learn to extract semantically meaningful information from the images themselves. The main difficulty in this approach is to avoid *representation collapse*, phenomenon where the networks ignore the inputs and produce identical and constant output vectors.

Recently, among the several existing ways to avoid model collapse, *distillation methods* have been pointed out as achieving better performance than others (Balestriero et al. (2023); Bardes et al. (2022)). In general, distillation methods train a student network to predict the representations of a teacher network. During the training phase, the gradients are only back-propagated through the student network, and the weights of the teacher are a running average of the weights of the student.

One of the most representative distillation SSL methods is the work of Caron et al. (2021). The authors designed an approach termed DINO as an acronym of "*knowledge **di**stillation with **no** labels*". DINO simplified SSL training by optimising the matching of the teacher network's output using a standard cross-entropy loss. Collapse prevention was achieved by including two simple operations in the teacher output, known as *centring* and *sharpening*. DINO could work on both transformer and convolutional architectures achieving SOTA accuracy on ImageNet. More interestingly, the trained encoders could obtain feature representations that explicitly contained a scene layout of the image, which could be used to generate accurate segmentation masks.

These promising models have also found their way into the field of medical imaging. Among the many papers applying SSL to medical images (Jiang et al. (2023); Kalapos and Gyires-Tóth (2023); Manna and Chakraborty (2022)), the work presented in Ye et al. (2022) deserves special attention. In this article, the authors proposed a DINO-based SSL method, called DeSD (**de**ep **s**elf-**d**istillation), which allowed the use of unlabelled data in the context of 3D medical image segmentation. In their model, both a student network and a momentum teacher were built by stacking several sub-encoders. The deep self-distillation supervision implied that the features of every student sub-encoder were optimised to match the teacher's output distribution. This technique resulted in superior pre-training of the segmentation network encoder compared to other existing SSL methods. When tested on seven downstream 3D medical segmentation datasets, their method outperformed training the same segmentation architecture from scratch and achieved state of the art results.

Considering the challenges associated with baseline NCCT AIS lesion segmentation, SSL can be identified as a promising approach to address the problem. The capacity of self-supervised learning techniques to make models extract semantically meaningful information coming from unlabelled datasets, represents a promising approach as a pre-training strategy in the context of AIS lesion segmentation. Given that SSL models have been shown to capture the scene layout of images, these methods may represent an unexplored way to exploit the intrinsic contextual information present in the brain NCCT image that is not necessarily related to the AIS lesion label.

In addition, these methods open the possibility of repurposing large amounts of unused, unlabelled NCCT images (with or without stroke) to improve segmentation performance. In a context of scarcity of good quality labelled AIS datasets, SSL could represent a more efficient use of the manual labelling process, limiting it to a subset of cases used for fine tuning to the downstream segmentation task and validating the results.

Lastly, U-shaped architectures and in particular nnU-Net, represent a standard segmentation baseline across many medical imaging modalities which has also been successful in the context of AIS lesion core segmentation. Therefore, the integration of SSL as a pre-training

strategy for nnU-Net encoder, in a similar manner to that done in Ye et al. (2022), may represent a synergistic way to integrate the aforementioned benefits of SSL with the highly successful self-configuring nnU-Net pipeline.

2.3. Contributions

1. A robust pre-processing pipeline for NCCT images which can work across many different datasets of variable quality.
2. A systematic pipeline to enhance nnU-Net autotuning framework with an additional encoder pretrainining in a self-supervised manner.
3. Use of a DeSD-like self supervised pre-training strategy to exploit large unlabelled NCCT (stroke positive and negative) datasets for the task of AIS segmentation.
4. Introduction of an asymmetry based data augmentation technique for achieving better latent representations for the context of stroke lesion segmentation on NCCT.
5. Pre-training nnU-Net's encoder with SSL is found to be an effective way for exploiting large amounts of unlabelled datasets, improving AIS final infarct segmentation performance on baseline NCCT images.

3. Material and methods

3.1. Datasets

In this work four datasets were utilised. A detailed description of each of them is presented below.

Acute Ischemic Stroke Dataset (AISD)

This dataset, published in Li et al. (2021), included cases of AIS with less than 24 hours from symptom onset to NCCT acquisition (n=397). For each case, NCCT, DWI and manual stroke lesion segmentation were provided. NCCT and DWI images were not registered. Patients underwent DWI within 24 hours of CT acquisition. Ground-truth labels were delineated on the NCCT by a physician using the MRI images as a reference. Important clinical information was missing from this dataset, such as the timing of DWI acquisition (pre/post endovascular treatment). As a result, it was not possible to determine whether the provided ground truth corresponded to final infarct or pre-treatment stroke core lesions. After visual inspection, five cases were discarded from the dataset due to large motion artefacts or non-overlapping ground truth with baseline imaging.

A Paired CT-MRI dataset for Ischemic Stroke Segmentation (APIS)

This dataset corresponded to the publicly released training subset of the ISBI 2023 APIS Challenge (Gómez et al. (2023)). The dataset (n=60) included patients over 18 years of age, collected from two Colombian clinics (FOSCAL and FOSUNAB), eight of whom were healthy controls. Each case included an NCCT image, the apparent diffusion coefficient (ADC) map derived from the DWI and a manual delineation of the stroke lesion core. No treatment was applied between NCCT and ADC (stroke lesion core as ground truth). Two neuroradiologists with more than five years of experience delineated the affected tissue over the DWI/ADC images. Eight cases were discarded due to high image corruption (missing slices, evident lesion mask misplacement), leaving a total count of forty-four stroke-positive cases.

In APIS dataset, NCCT and ADC maps were provided registered and skull-stripped. However, after a visual inspection, the registration process was found suboptimal. This had two negative consequences: non-brain structures (e.g. bone) were visible in the skull-stripped image and the labels were misregistered.

icometrix Acute Ischemic Stroke Dataset (icoAIS).

This was an in-house private set of cases provided by the Klinikum rechts der Isar (Munich, Germany). The collection included acute and early subacute stroke patients (n=159) and healthy patients (n=8), all over 18 years of age. Three images were available for each case: NCCT, DWI and ADC map. The MRI images were provided already skull-stripped and registered to a common space. MRI images were acquired after successful revascularisation therapy. The collection included a wide range of infarct patterns in all vascular territories, even including posterior circulation infarcts, which are common in clinical practice but not commonly studied in the literature.

Ground truth labels for icoAIS were obtained in two different ways. In half of the cases (n=79), the voxel-level labels involved a high-quality hybrid human-algorithm annotation process described in Hernandez Petzsche et al. (2022). Since the process involved neuroradiologists with more than ten years of experience reviewing the MRI images, this subset was referred to as *gold standard* labels. For the remaining stroke-positive cases (n=80), *silver standard* labels were obtained by running SEALS -the publicly available[1] ISLES22 winning solution- over the DWI images. A stroke expert reviewed the annotations and found them to be adequate and highly correlated with the available gold standard annotations.

Collaborative European Neuro-Trauma Effectiveness Research in Traumatic Brain Injury Dataset (CENTER-TBI)

This collection of NCCT images was a multi-centre, multi-scanner dataset presented in Maas et al. (2015). From the complete dataset, only a selection of NCCT scans identified by expert review as not having abnormal TBI-related findings was kept (n=637). This re-

[1] https://github.com/Tabrisrei/ISLES22_SEALS

sulted in a very diverse dataset of healthy (non-stroke) patients.

In all cases, due to the retrospective nature of this work and the rigorous anonymisation of the data, it was not necessary to obtain informed consent from the patients. In the particular case of APIS, the data was used in agreement with the APIS challenge signed informed consent.

3.1.1. Dataset partitioning

The complete dataset was split into three subsets: training, validation and test (70-20-10% respectively). The partitioning was done individually for each dataset at the patient level. For stroke-positive datasets, the partitioning was done stratified by lesion location and size. This ensured that all subsets had equal representation of right, left and bilateral lesions and lesion sizes. In the specific case of icoAIS the dataset was also stratified by the labelling standard (gold vs. silver).

3.2. Preprocessing

As a first pre-processing step, all the images were turned into NIfTI format. Cases that where acquired in "tilted" fashion were "un-tilted" during the DICOM to NIfTI conversion using the NITRC conversion tool (Li et al. (2016)).

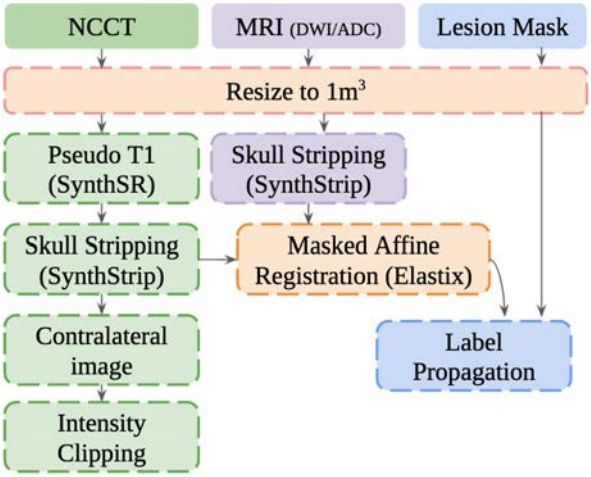


Figure 1: Preprocessing pipeline

It is noteworthy that there were significant differences between the datasets used. Diversity is desirable in the context of SSL, but to reduce the risk of the algorithms exploiting meaningless shortcuts or biases (i.e. distinguishing data origins by their skull-stripping quality), the datasets were homogenised as much as possible. To do this, a robust preprocessing pipeline (summarised in Figure 1) was applied to all datasets. It comprised six steps:

a. Resampling. All volumes were resampled to $1mm^3$ resolution using a linear interpolation for NCCT, ADC and DWI images and a nearest neighbour interpolation for the ground truth mask.

b. Skull Stripping. NCCT and DWI/ADC images are not characterised by having high contrast differences between the different brain soft tissues. As a consequence, popular skull stripping methods (Ashburner and Friston (2005); Isensee et al. (2019); Lutkenhoff et al. (2014)) mainly developed to work on high resolution T1 MRI images gave poor results when applied to the desired modalities. In addition, for APIS, all methods failed due to the pre-existing sub-optimal preprocessing. Robust results were obtained by combining two models from the publicly available FreeSurfer toolbox: SynthSR and SynthStrip (Hoopes et al. (2022); Iglesias et al. (2023)).

In both cases, the respective authors used a clever synthetic data generation technique to obtain robust models across multiple resolutions and contrasts. The authors show that SynthSR is able to generate a high-resolution T1 MRI out of any brain MRI image and has a reasonable performance when using CT scans. SynthStrip is a brain segmentation model that is very robust across different brain MRI modalities, but it did not work very well when applied directly to the NCCT image. Instead, generating a pseudo-T1 first and applying SynthStrip over it gave the best results. SynthStrip was applied directly to ADC and DWI MRI images with good results.

c. Registration. For APIS and icoAIS, a brain-masked affine registration of the MRI image to the NCCT space was performed using the Elastix toolbox (Shamonin (2013)). This involved a pyramidal registration with mutual information as the objective function. After this, the stroke lesion mask was propagated to the NCCT space using the same transformation but with nearest neighbour interpolation.

d. Contralateral image. To obtain the mirrored brain with respect to the inter-hemispheric plane, the NCCT was first registered to an NCCT MNI space template (Rorden et al. (2012)) using brain-masked affine registration. A left-right flip was then performed and the resulting image was masked-affine registered to the original NCCT image. In this way, the desired image ended in the original patient space and the effect of gross normal asymmetries in brain shape was reduced.

e. Intensity clipping. Following the literature and the recommendations done by an expert in stroke imaging, the NCCT and the contralateral NCCT images were intensity clipped to the range [-100, 400], leaving unchanged the range in which both brain soft tissue and stroke lesions have their intensities.

3.3. Method overview

As mentioned in the introduction, this work uses a variant of the two-step SSL paradigm presented in Ye et al. (2022). Unlike the cited work, the nnU-Net self-configuring model is used as the base architecture and

is enhanced by adding SSL pre-training to its encoder part.

Introduced in the work of Isensee et al. (2020), nnU-Net is a self-configuring method for deep-learning biomedical image segmentation. In terms of implementation, it consists of a robust pipeline with two stages: first, it finds the best configuration of the UNet model for any new dataset, and then, based on the conclusions of this step, the tailored training can be performed. In the first step, the data pre-processing, network architecture, training details and post-processing stages are decided. Two core elements are involved in this process: a *dataset fingerprint* and a *pipeline fingerprint*. The first one is a standardised dataset representation comprising key properties such as the image size, voxel spacing and class ratios. The second one registers a collection of choices made during the automatic optimal method design (i.e. batch size, patch size, network topology, etc.) and serves as a recipe followed during training to instantiate the model and the training machinery.

Given the success of this self-tuned pipeline, we used it as a basis for a four steps training strategy:

1. nnU-Net configuration according to the dataset used for SSL.
2. Self-Supervised pre-training of nnU-Net encoder.
3. nnU-Net configuration according to the dataset used for supervised learning.
4. Transfer learning and supervised training of nnU-Net segmentation network.

In the first step, the objective was to use the optimal architecture configuration and data pre-processing of nnU-Net, but for self-supervised training. To this end, an nnU-Net dataset containing the full set of NCCT images (stroke and healthy cases) was generated and the nnU-Net self-configuration pipeline was run over it. By default, during the dataset fingerprint extraction, nnU-Net computes the intensity statistics -that are later used for normalising the images- from foreground region defined by the segmentation mask. To be able to process healthy cases and to obtain a more general normalisation strategy, nnU-Net pipeline was fed with the brain masks as if they were the segmentation targets, so that all the whole brain region was treated as foreground. Three results were kept from this step: the pre-processed images, the dataset fingerprint and the pipeline fingerprint.

In the second step, the pipeline fingerprint was used to instantiate the complete UNet model, discarding the decoder part and adding the additional modules required for deep self-distillation training (see details in Section 3.4). This distillation mechanism was implemented to dynamically adapt to the number of encoding steps defined by the nnU-Net pipeline. Finally, the encoder was trained in SSL fashion and its weights were re-adjusted (removing the extra SSL modules) to the original nnU-Net encoder structure.

In the third step, the self-configuring nnU-Net pipeline was run a second time. This time, only the subset of desired labelled images were included and the stroke lesion masks were given as segmentation targets. The resulting dataset fingerprint was modified by replacing the intensity statistics with those from the full NCCT dataset, so that the pre-processed inputs were in the same intensity space used to pre-train the encoder. The preprocessing was then run and the pipeline fingerprint was generated.

Finally, the nnU-Net supervised training pipeline was run using the pre-trained weights of the encoder as a starting checkpoint.

3.4. Self-supervised learning details

The SSL method implemented in this work shared the same overall structure with DeSD (Ye et al. (2022)). Its general outline can be appreciated in Figure 2. The overall strategy was based on knowledge distillation, training a student network g_{θ_s} to match the output of a teacher network g_{θ_t} (where θ_s and θ_t were their parameters respectively).

Both networks roughly shared the same architecture, but the student one was decoupled into N sub-encoders: $g^i_{\theta_s}$ for i from 1 to N. The network g_{θ_s} was generated by adding a projection head at the end of each of the 6 encoding stages determined by the nnU-Net configuration pipeline. Naming the projection head h, each stage of nnU-Net encoder f^i ($i = 1, ..., N$) and the complete encoder f, the overall network could be formally written as: $g^i_s = h^i \circ f^i_s$. The same holds for the teacher network, but using only the complete encoder: $g_t = h_t \circ f_t$.

In all cases, the projection head (h) consisted of a multi-layer perceptron (MLP) of three layers plus a final weight-normalised fully connected layer of dimension K. The two MLP hidden layers were of dimension 2048, with batch normalisation and Gaussian Error Linear Units (GELU) activation function. The MLP output had a dimension of 256, with no batch normalisation or non-linearity applied. In summary, the complete projection head mapped the output of each stage of the student network and the final teacher network into a K dimensional representation, with $K = 65536$ in our case.

As done in Caron et al. (2021), after the K dimensional teacher representations were obtained, a centring was applied to avoid model collapse. The centring depended on the first order batch statistics and can be interpreted as adding a bias term c to the teacher: $g_t(x) \leftarrow g_t(x) + c$. The centre c was updated with an exponential moving average rule:

$$c \leftarrow mc + (1 - m)\frac{1}{B}\sum_{i=1}^{B} g_{\theta_t}(x_i), \tag{1}$$

where $m = 0.9$ and B is the batch size.

Given an input image x, the resulting representations from the N=6 student sub-encoders and the teacher one

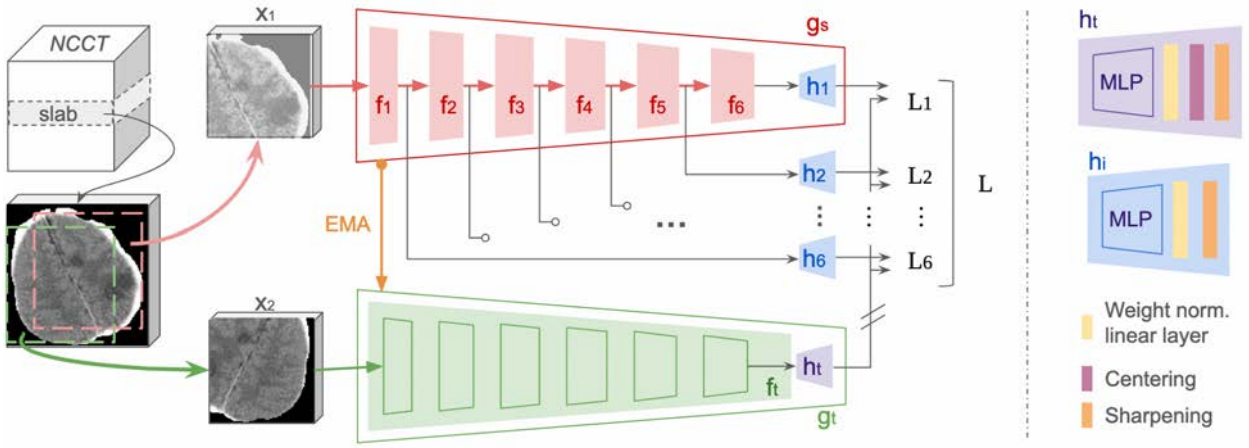


Figure 2: Self-supervised training method, see Section 3.4 for details. Subindex *s* and *t* stands for student and teacher respectively, *L* for loss, *EMA* for exponential moving average and *MLP* for multi-layer perceptron.

were transformed into probability distributions over the K dimensions (denoted P_s^i for $i = 1, ..., N$ and P_t respectively). Each probability P was obtained by normalising the output of the corresponding network g with a softmax function:

$$P(x)^{(j)} = \frac{exp(g(x)^{(j)}/T}{\sum_{k=1}^{K} exp(g(x)^{(k)}/T}, \tag{2}$$

where $j = 1, ..., K$ and $T > 0$ was a temperature parameter that, as indicated in Caron et al. (2021), had a sharpening effect that reduced the possibility of representation collapse.

As usually done in SSL, from a given image, a set $V = \{x_1, x_2\}$ was generated with two differently distorted views (crops) of it. Then, x_1 and x_2 were passed through both the student and the teacher networks, in order to obtain their respective embeddings.

During training, only the student weights were updated through gradient back-propagation. The training objective was to match each student embedding with the teacher's one by minimising the cross-entropy loss:

$$\min_{\theta_s} \frac{1}{N} \sum_{i=1}^{N} \sum_{x \in V} \sum_{\substack{x' \in V \\ x' \neq V}} H(P_t(x), P_s(x')), \tag{3}$$

where $H(a, b) = -a \log b$.

The weights of the teacher network were updated using an exponential moving average of the student ones (momentum encoder). The updating rule was defined by: $\theta_t \leftarrow \lambda\theta_t + (1 - \lambda)\theta_s$. Where λ followed a cosine schedule from 0.9996 to 1 during training.

Finally, once the model was trained, the features used in downstream task are the ones from the backbone f_t, dropping the projection head.

3.4.1. Distorted views strategy

The SSL literature suggests that inputs to Siamese networks should follow two recommendations: The view size should contain more than 50% of the original image, and that the larger the batch size used during training, the better. However, as the proposed models had 3D inputs, which increased their GPU memory consumption, a trade-off was made between these two requirements. An anisotropic patch size of 112x112x16 (favouring inter-hemispheric contextual content) and a batch size of 64 (the largest the memory would hold) were chosen.

Given that the median image size in our datasets was 169x138x139, to avoid sampling two completely different views from the volume, an online patch sampling strategy of two stages was used. First, from the region defined by the brain's bounding box, a slab of 24 slices was sampled along the z axis and then the two patches x_1 and x_2 were sampled from within the slab. This ensured a high degree of overlap between the paired views. In addition, to prioritise slabs with higher brain content, if the sampled slab contained less than 40% of brain, it was replaced by a new sample with a probability equal to the background content percentage (the less brain content, the more chances to take another sample).

Several data augmentations were applied to both views: flipping, scaling, Gaussian noise, Gaussian blur, gamma intensity transformation, change of image brightness and contrast. These were the same ones used in Ye et al. (2022) but applied on a volume fashion and not slice-wise.

3.4.2. Training and evaluation details

Evaluating the progress of SSL methods is not a simple task. Common ways to evaluate the quality of the obtained representations are linear probing with a classification task, k-nearest neighbours clustering or directly training over the downstream task at each epoch. In our case, training the decoder stage of the segmentation network at each epoch was computationally prohibitive. For this reason, the training dynamics were

evaluated by checking the training and validation loss curves and using *RankMe*.

RankMe is a recently proposed metric to evaluate SSL training performance in a fully unsupervised manner (Garrido et al. (2023)). It represents an indirect way to evaluate the information content of a set of representations by assessing an approximation to their rank. The higher the RankMe, the more linearly independent the representation components are and more the variance of the data is distributed among them, showing an empirical positive correlation with their performance in different downstream tasks. Following the cited publication, RankMe was computed over the latent representation of 256 elements produced by the MLP of the teacher's projection head.

Preliminary experiments were carried out to become familiar with the learning dynamics of the models and to find a suitable set of hyperparameters and training choices.

An AdamW optimiser was used with a batch size of 64. The learning rate was linearly ramped up during the first 10 training epochs to its base value of 0.1. After this warm-up the learning rate was decayed with a cosine schedule. Weight decay also followed a cosine schedule from 0.04 to 10^{-5}. The temperature T_s was set to 0.1 while a linear warm-up from 0.04 to 0.07 during the first 10 epochs was used for T_t. In order to reduce memory consumption 16-bit floating precision was used for the model weights. Independently on the dataset used, in all the experiments SSL was trained for 100 epochs after which train loss, validation losses and RankMe curves reached a plateau. Each epoch consisted of 9600 iterations. Since there were no clear signs of overfitting or model performance decay (or improvement) at the final plateau, the model from the last epoch was kept as the pre-trained checkpoint.

3.4.3. Latent representation projection and model interpretability

After training the encoder in the SSL fashion, two techniques were used to interpret the obtained representations: attribution maps and t-SNE dimensional reduction to observe clustered patterns.

Attribution maps were computed using Integrated Gradients method (Sundararajan et al. (2017)). Since our network did not output class wise predictions, a sum of all the elements in the final representation was added at the end of the network before applying the attribution method. The resulting maps highlighted the voxels whose change most affected the entire latent representation.

To analyse the projections in a more systematic way, a t-distributed Stochastic Neighbour Embedding (t-SNE) projection of the teacher representations onto a two-dimensional space was obtained (van der Maaten and Hinton (2008)). Given the resulting clustered nature of the resulting 2D space, a set of representative points of the visible groupings were selected and their six nearest neighbours were obtained. For these 6 cases, axial slices of the NCCT volumes and their attribution maps were plotted with the aim of detecting common salient features used by the network to generate the representations (see Figure 8 for an example).

3.5. Supervised Segmentation details

Among the different experiments performed in the supervised segmentation training, the main objective was to compare the impact of the different training strategies on the performance of the final infarct segmentation on baseline NCCT images. In this sense, we decided to focus on the icoAIS dataset for supervised segmentation. This was the best described dataset, the one with the best image quality and the only one that was certain to contain only final infarct ground truth.

As a first step, several experiments were performed to train the nnU-Net model from scratch with the icoAIS dataset to gain insight into the training dynamics. From these experiments it was found that the 3D nnU-Net architecture outperformed the 2D version and therefore the former was retained throughout the work.

After running the nnU-Net self-configuring pipeline over the dataset used for supervised learning, the resulting model architecture is illustrated in Figure 3. The model details follow the overall rules defined in Isensee et al. (2020). However, the main design specifications are presented below.

The model consisted of six encoding and six decoding stages 3D U-shaped architecture, using only plain convolutions, instance normalisation and Leaky ReLU non-linear activation function. Each resolution stage of both the encoder and the decoder consisted of two computational blocks of convolution-normalisation-activation function. Down-sampling was done with strided convolutions and up-sampling with transposed ones. The initial number of feature maps was set to 32 and doubled (halved) with each down-sampling (up-sampling) operation, the number of feature maps across the network was capped at 320 to limit the chance of overfitting.

During training, each epoch implied 250 iterations, where the minibatch size was 2. Stochastic gradient descent with Nesterov momentum ($\mu = 0.99$) and an initial learning rate of 0.01 was used to optimise the network weights. The learning rate was decayed through the training according to the 'poly' learning rate policy, $(1 - epoch/epoch_{max})^{0.9}$. The loss function was the average of binary cross-entropy and soft dice losses.

The network was trained with deep supervision, where additional losses are added in the decoder at all but the two lowest resolutions, each using a down sampled version of the ground truth mask. The training objective was the sum of the losses at all resolutions, $L = w_1 xL_1 + ... + w_4 xL_4$, where the weights $w_i = \frac{1}{2^i} w_1$ were later normalised to sum to 1.

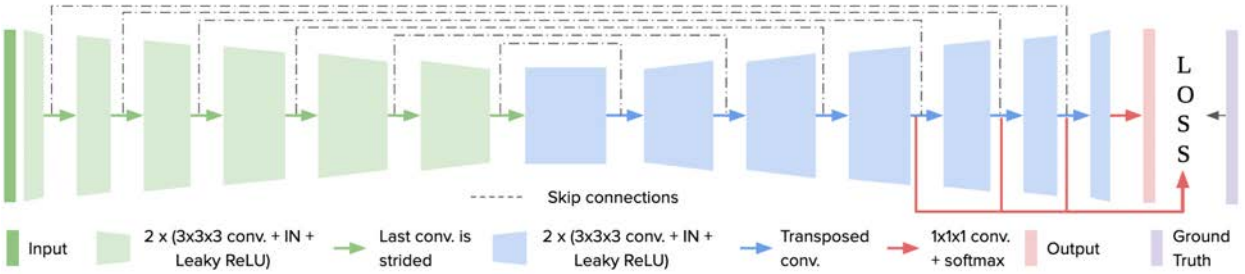


Figure 3: Alternative representation of the U-Net model, highlighting the encoding-decoding nature of the architecture

Since the resampling was performed as part of our preprocessing pipeline, the nnU-Net preprocessing consisted only of normalising the images using the foreground voxels statistics. The 0.5 and 99.5 percentiles were used for clipping, and the post-calculated mean and standard deviation were used for z-score normalisation. Samples for the mini-batches were selected from random training cases. Class unbalance was handled with over sampling by forcing that at least half of the samples in the minibatch had to contain stroke on it. The patch size used was 160x160x96. A variety of data augmentations were applied on the fly during training: rotations, scaling, Gaussian noise, Gaussian blur, brightness, contrast, low resolution simulation, gamma correction and mirroring.

Unless otherwise stated, all networks were trained for 100 epochs. After these epochs, the model trained from scratch on the icoAIS dataset reached a plateau on the validation exponential moving average (EMA) pseudo-dice, and the validation and train loss curves began to show overfitting patterns (the former stopped decreasing, while the latter continued to decrease). In all cases, the models were trained on the training set, model selection was done with the validation set, and the test set was left aside in both procedures.

The nnU-Net pipeline retains the best model from all training epochs as the one that maximises the pseudo-dice EMA over the validation set. This dice approximation is computed by considering each batch of samples from the validation cases as a case itself. This strategy was shown in the original paper to be a good compromise between computational cost and performance impact. However, in the problem addressed in this work, it resulted in a very noisy validation pseudo-dice curve, which - even when smoothed by the EMA - affected the selection of the best model. In the preliminary experiments, it was observed that when the model was trained for 300 epochs, the best selected model did not achieve the highest performance over the full images in the validation set. This suggested that, in our specific problem, nnU-Net tended to select slightly overfitted models as best, which was "manually" prevented by limiting training to 100 epochs.

Another preliminary finding was that when nnU-Net supervised training was repeated on the same cases and experimental design, but with different random initialisations, there was significant variability in performance on the full validation cases. In an attempt to minimise this confounding noise, three different runs with different random initialisations were made for each supervised experiment. For each of them, the best model was selected according to the nnU-Net criteria and the majority voting ensemble was computed from the predictions of these three models.

Finally, during inference time, images were predicted using a sliding window approach with a window size of 96x160x160 and a stride of 48. Gaussian importance weighting was applied, increasing the weight of central voxels in the softmax aggregation. Test time data augmentation was applied by mirroring all axes.

3.5.1. Training nnU-Net with pretrained weights

When performing transfer learning from a pre-trained SSL model to a downstream task, two main strategies can be used to prevent the encoder from "forgetting" what it has learned: freezing the weights of the pre-trained encoder (for a fraction of epochs or for all epochs) or using smaller learning rates for the encoder. In order not to modify the nnU-Net pipeline too much, we decided to go for the freezing strategy. Four strategies were briefly explored: leaving all parameters unfrozen, unfreezing the encoder after 33 epochs or after 66 epochs, and leaving the encoder frozen for the entire training procedure. The best resulting strategy was to leave all parameters unfrozen and was therefore used in all experiments presented in this work.

3.6. Metrics

The performance of the models in the segmentation task was evaluated using a set of different metrics. Among the metrics commonly reported in the medical imaging community, the Dice Score Coefficient (DSC) and the 95% Hausdorff Distance (HD95) were used. The former evaluates the voxel overlap between the segmentation and the ground truth, ignoring the true negatives in the background, however Dice is not well suited to detecting outliers in the contour prediction. HD compares the greatest distance between predicted

and ground truth contours. For both the DSC and HD95, calculated at subject level, we report the mean, median and interquartile range

From a clinical perspective, since lesion volume plays a key role in patient selection criteria, it is important to measure how close the predicted volume was to the ground truth. In this sense, three measures were included: absolute volume difference, volumetric Spearman correlation and interclass correlation coefficient (two-way mixed effects, single rater consistency definition (ICC(3,1)) (Koo and Li (2016)). All metrics were calculated at the subject level.

Since the ground truths are derived from post-treatment DWI images, minor embolic lesions of no clinical significance and characterised by volumes $< 3mL$ may be included in the ground truth whereas they were not present in the baseline image. Following Giancardo et al. (2023) and the suggestions from a stroke imaging expert, we decided to report the metrics for two different scenarios. One where all lesions are retained in the ground truth and a second one, where only lesions larger than 3 mL are retained.

For this two scenarios both the results on the independent validation and test sets are reported in the results section.

3.6.1. Statistical Analysis

Dice scores across experiments were statistically compared using the Wilcoxon signed-rank test after rejecting normality with the Shapiro-Wilk test and QQ plot analysis.

3.7. Experiments

3.7.1. Baselines

The effect of using an SSL pre-training strategy was compared against two baselines:

- Training the nnU-Net model from scratch using only the icoAIS dataset, run hereafter referred to as **FS-STKi**, following the convention (training mechanism)-(supervised dataset), where *FS* stands for From Scratch, *STK* for stroke and *i* for the particular icoAIS dataset.

- Training the nnU-Net model from scratch using all the available labelled data coming from the three stroke positive datasets: AISD, APIS, icoAIS. Similar as before, this run is referred to as **FS-STKp**, where the *p* stands for positive.

It is important to note that for the second baseline, the training was conducted for 300 epochs. Because of the increase in the training data, the training convergence took longer. At around 300 epochs, the same rationale mentioned for selecting 100 epochs was fulfilled.

3.7.2. Exploring different datasets for SSL pre-training

In order to evaluate the benefits of using the pre-trained encoder, several experiments were carried out. The first one involved studying the influence of using different datasets during SSL pre-training. Three scenarios were investigated:

- Training on all the available NCCT images (abbreviated as *"ALL"*). This involved using the AISD, APIS, icoAIS and CENTER-TBI datasets, with the hypothesis that including the greater diversity of images in the dataset would allow the model to learn better representations.

- Training with all stroke-positive NCCT images (*STKp*). This involved using the AISD, APIS, icoAIS datasets, hypothesising that including of only stroke-positive images might allow the model to somehow capture better suited representations for the problem under assessment.

- Training only on healthy/non-stroke patients (*STKn*). This implied using only the CENTER-TBI dataset, with the idea that pre-training on healthy patients and fine-tuning on stroke-positive cases might allow the model somehow exploit the difference (presence of lesions) between these image sets.

For all these different datasets, the SSL pre-training was done as described in Subsection 3.4.2. Once the encoder was pre-trained, the supervised training was done *using only the icoAIS dataset*. In the following, these supervised pre-trained experiments are identified respectively as **ALL-STKi**, **STKp-STKi**, **STKn-STKi**, following a convention close to the one defined above (SSL dataset)-(supervised dataset).

3.7.3. Symmetry focused data augmentation technique

In self-supervised learning the data augmentation techniques used to generate the two different views from the same image play an important role in the representations learned by the model. In an attempt to integrate the inter-hemispheric asymmetries more explicitly into the SSL pre-training pipeline, a specific data augmentation technique was developed. In detail, when a patch was sampled from the original NCCT volume, the same patch location was sampled from the contralateral image and both views were later subjected to the regular data augmentation techniques.

The idea behind this experiment was that with this augmentation, the resulting model representations would be more agnostic to brain asymmetries. Derived from this, training the encoders with healthy patients only, the model would disregard normal asymmetries, which could have an impact when trained in a supervised manner with the presence of stroke-induced asymmetries.

To explore this idea, we pre-trained the encoder with the three dataset configurations presented in the previous section, but added the symmetry augmentation with a probability of 0.7 each time an image was sampled (regardless of its stroke content).

Once the encoders were pretrained, the complete supervised nnU-Net was trained using only the icoAIS dataset. These experiments are identified as **ALL-STKi-SA**, **STKp-STKi-SA**, **STKn-STKi-SA**, following a close convention to the one defined above (SSL pretraining dataset)-(supervised dataset)-(symmetry augmentation).

3.7.4. Additional experiments

An early phase of this work involved the participation in the ISBI 2023 APIS Challenge. Following the reported benefits of including inter-hemispheric differences in ischemic stroke segmentation models, a 3D nnU-Net was trained over the complete stroke-positive training set (*STKp*) with the same specifications as above, but using a different input. The NCCT and the difference from its sagittally mirrored version were used as a multi-channel input to the model. This model achieved first place in the competition by a wide margin over the other participants (for further details see Appendix A).

With the aim of combining this initial achievement with the main line of research of this work, a further set of experiments was carried out using the multi-channel 3D input to train both the encoder with SSL and the complete nnU-Net in a supervised fashion (with or without pre-training). For the sake of clarity, the details and results of these side experiments are included in Appendix B.

3.8. Computational resources

All the models were implemented using Python 3.10.9 and PyTorch 2.0. All the experiments were run on a 64-bit GNU/Linux (Ubuntu 20.04) server with an 8-core AMD EPYC 7R32 CPU (2.8 GHz) with 32 GB of RAM and a single NVIDIA A10G GPU card with 24 GB GDDR6 of memory using CUDA 11.6.

4. Results

4.1. Results over icoAIS validation set

In Tables 1 and 2 the quantitative segmentation results obtained on the icoAIS validation set are presented. The first table shows the results considering all lesion sizes and the second one only considering lesions bigger than 3 mL.

Firstly, from the two result summaries, focusing on DSC as usually done in the literature, we can see that pre-training the U-shape model's encoder with SSL gave an improvement in performance compared to training from scratch. Almost all the methods pre-trained with SSL had a higher mean and median Dice than *FS-STKi*. Secondly, when considering the influence of the dataset used in SSL pre-training, using all the available NCCTs (*ALL*) was on par to using only the stroke positive datasets (*STKp*). However, both achieved superior Dice scores than using only stroke-negative (healthy) data (*STKn*). Finally, when evaluating the use of the ad hoc symmetry augmentation technique, it can be seen that it slightly improved the performance when pre-training with all the NCCTs or only the stroke-positive datasets.

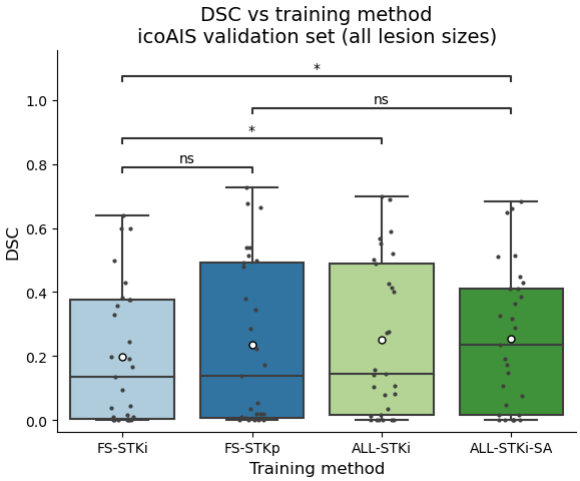


Figure 4: Dice Score Coefficient for the best performing training methods computed over the *icoAIS validation set* considering all lesion sizes. The statistical significance were determined with a paired Wilcoxon rank test, where *ns* indicates $0.05 < p <= 1$ and * indicates $0.01 < p <= 0.05$

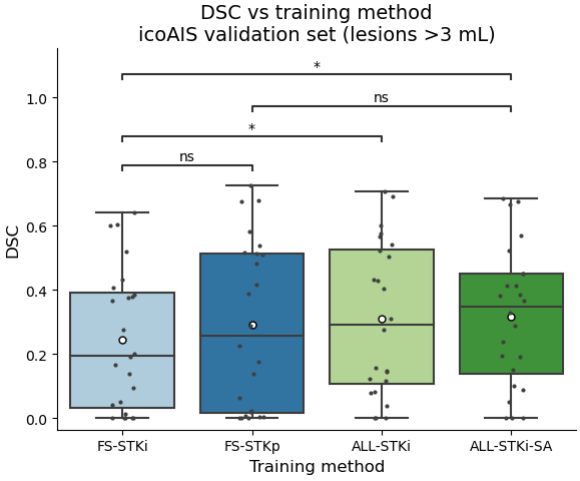


Figure 5: Dice Score Coefficient for the best performing training methods computed over the *AIS validation set* considering lesions $> 3mL$. The statistical significance were determined with a paired Wilcoxon rank test, where *ns* indicates $0.05 < p <= 1$ and * indicates $0.01 < p <= 0.05$

Again focusing on Dice, from all the SSL pre-training variants, the best performance was achieved when using all available NCCTs with the symmetric data augmen-

Table 1: Performance measures on *icoAIS validation set* considering all lesion sizes (cases n=29).

	DSC ↑		HD95 [mm] ↓		AVD [mL] ↓		Corr ↑	ICC ↑
Experiment	Mean	Median(Iqr)	Mean	Median(Iqr)	Mean	Median(Iqr)		
FS-STKi	0.1981	0.135 (0.373)	45.19	47.27 (14.48)	**19.74**	9.51 (17.03)	0.52	0.70
FS-STKp	0.2361	0.138 (0.487)	**39.39**	**35.52** (23.84)	21.83	10.23 (27.23)	0.65	**0.74**
ALL-STKi	0.2510	0.144 (0.473)	45.72	50.32 (32.85)	21.56	**8.61** (21.01)	0.65	0.63
ALL-STKi-SA	**0.2554**	**0.237** (0.395)	44.33	47.69 (33.45)	21.84	9.23 (19.76)	**0.70**	0.63
STKp-STKi	0.2503	0.166 (0.431)	49.94	50.47 (29.35)	20.29	9.74 (17.47)	0.60	0.67
STKp-STKi-SA	0.2525	0.186 (0.424)	43.76	48.19 (30.80)	20.27	9.55 (17.70)	0.61	0.65
STKn-STKi	0.2195	0.133 (0.396)	52.43	48.28 (40.01)	24.23	11.22 (20.41)	0.58	0.61
STKn-STKi-SA	0.2096	0.083 (0.336)	49.86	47.93 (38.67)	26.54	12.73 (25.60)	0.64	0.54

Table 2: Performance measures on *icoAIS validation set* considering lesions > 3*mL* (cases n=24).

	DSC ↑		HD95 [mm] ↓		AVD [mL] ↓		Corr ↑	ICC ↑
Experiment	Mean	Median(Iqr)	Mean	Median(Iqr)	Mean	Median(Iqr)		
FS-STKi	0.2451	0.1958 (0.358)	44.37	43.60(26.20)	**21.18**	9.91 (19.86)	0.50	0.7
FS-STKp	0.2720	0.2981 (0.405)	44.87	41.75(38.07)	27.99	14.57 (34.48)	0.69	**0.73**
ALL-STKi	0.3101	0.2926 (0.420)	43.12	39.16(32.38)	23.40	**7.00** (24.26)	0.65	0.62
ALL-STKi-SA	**0.3171**	**0.3477** (0.312)	**38.52**	**34.70**(37.78)	24.01	11.50 (20.73)	**0.71**	0.63
STKp-STKi	0.3103	0.3307 (0.358)	51.64	45.91(41.46)	21.07	8.91 (17.02)	0.69	0.66
STKp-STKi-SA	0.3122	0.3589 (0.376)	44.74	45.16(35.71)	21.32	7.61 (20.05)	0.66	0.65
STKn-STKi	0.2725	0.2705 (0.412)	47.76	41.04(44.06)	27.10	13.07 (18.84)	0.62	0.61
STKn-STKi-SA	0.2608	0.1986 (0.369)	45.92	38.59(34.48)	29.95	15.91 (35.32)	0.65	0.52

tation (*ALL-STKi-SA*). See Figure 4 for a comparison of Dice for the best performing methods. Considering all lesion sizes, a mean Dice of 0.2554 ± 0.225 and a median Dice of 0.237 ± 0.395 were obtained, which were significantly higher than those obtained for training from scratch only with icoAIS data (mean DSC of 0.1981 ± 0.214 and median of 0.135 ± 0.373).

More interestingly, the results obtained with SSL pre-training on all the data (*ALL-STKi-SA*) were superior to those obtained when the supervised model was trained from scratch with all labelled datasets (*FS-STKp*) (mean DSC 0.2361 ± 0.255 and median 0.138 ± 0.487).

When considering the results obtained by including only lesions larger than 3 mL, two things must be noted. The results stated in the previous paragraph still hold in this case -as can be seen in Table 2- with *ALL-STKi-SA* SSL pre-trained model outperformed its counterpart trained from scratch (see Figure 5 for a boxplot comparison of the best methods).

For the other quantitative measures presented in Tables 1 and 2, the results were not as clear as in the case of DSC. In the case of the 95% Hausdorff distance, regardless of the lesion size considered, *ALL-STKi-SA* slightly outperformed training from scratch with the same supervised learning dataset. However, this was not the case when using all labelled cases, where the SSL model

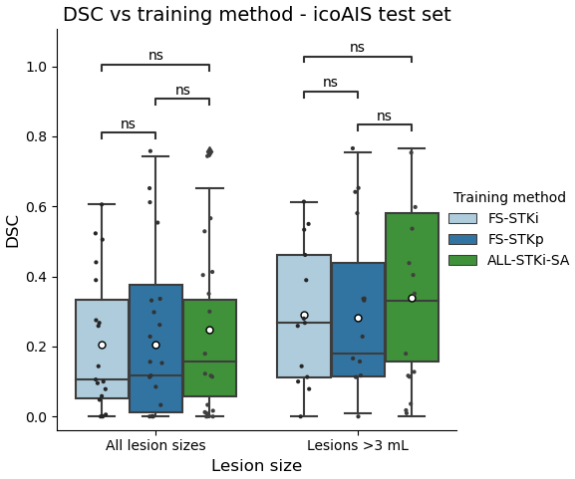


Figure 6: Dice Score Coefficient for the best performing training methods computed over the *icoAIS test set* ablated by lesion size. The statistical significance were determined with a paired Wilcoxon rank test, where *ns* indicates $0.05 < p <= 1$

was superior only when excluding the small volumes. When considering the absolute volume difference, the results are even more dispersed, indicating that the best performing method across all runs was *FS-STKi* if we consider the mean value, or *ALL-STKi* if we consider

Table 3: Performance measures on **icoAIS test set** for both lesion sizes criteria (cases n=19 for *All* and n=14 for > 3*mL*).

	DSC ↑		HD95 [mm] ↓		AVD [mL] ↓		Lesion size
Experiment	Mean	Median(Iqr)	Mean	Median(Iqr)	Mean	Median(Iqr)	
FS-STKi	0.2053	0.106 (0.279)	46.99	48.75 (19.71)	26.81	8.92 (30.37)	All
FS-STKp	0.2059	0.117 (0.366)	**45.23**	**42.40** (33.66)	23.77	11.28 (28.79)	All
ALL-STKi-SA	**0.2468**	**0.156** (0.275)	45.81	42.56 (27.89)	**21.51**	**6.70** (21.01)	All
FS-STKi	0.2915	0.268 (0.349)	42.91	41.27 (28.61)	30.43	5.47 (50.83)	> 3*mL*
FS-STKp	0.2833	0.180 (0.325)	**36.69**	**37.47** (38.61)	**22.24**	9.53 (31.37)	> 3*mL*
ALL-STKi-SA	**0.3406**	**0.332** (0.424)	42.66	42.56 (41.39)	23.27	**5.32** (29.38)	> 3*mL*

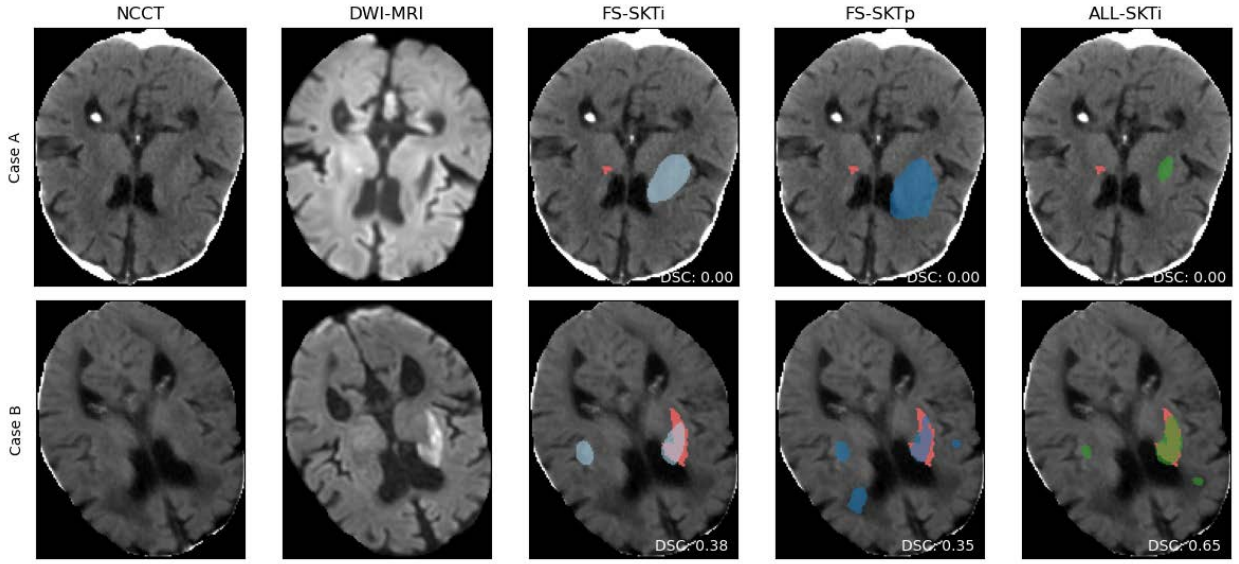


Figure 7: Example results from the **icoAIS test set**. In the upper row a case (A) in which all the methods failed, in the bottom row a case (B) in which all the methods had reasonable performance. In all the cases the GT is shown in red with the model prediction overlaid in colours different from red

the median. Finally, the model selected as best by the Dice criteria showed the highest Spearman volume correlation, and all the models showed a moderate ICC, both when including and excluding lesions smaller than 3mL.

4.2. Results over icoAIS test set

In Table 3 the quantitative segmentation results obtained on the icoAIS test set are presented. The upper part shows the results considering all lesion sizes (n=19) and the lower part only considering lesions bigger than 3 mL (n=14). In both cases, only the best performing SSL pre-trained method according to the validation set was compared against the two baselines.

When considering Dice score as the main metric (results summarised in Figure 6), although the differences were not statistically significant, the same trend as seen in the validation sets could be observed. Pre-training the nnU-Net encoder with all the available NCCTs in the SSL fashion, and then fine-tuning with only the cases from the icoAIS dataset, resulted in a better performance (median DSC for all lesion sizes: 0.156, and for lesions > 3*mL*: 0.332) than training from scratch (median DSC for all lesion sizes: 0.106, and for lesions > 3*mL*: 0.268), and even slightly better than training from scratch with all the labelled data. Although the SSL pre-trained model did not achieve the best HD95, it slightly outperformed its counterpart trained from scratch, and it achieved the best average volume difference of the three models considered. Spearman correlation and ICC values were omitted due to the low confidence in their results given the small size of the test set.

Figure 7 shows qualitative results for two cases from the test partition of the icoAIS dataset. The top row persents a challenging case that none of the methods could segment correctly, and the bottom row shows a case where all methods performed well. In both cases, it is important to note that the SSL pre-trained model (last column) was more specific than its counterparts trained

from scratch.

4.3. Interpreting the SSL pre-trained encoder

Figure 8, shows the results of applying some techniques to interpret the latent representations obtained with the best SSL pre-trained encoder. The top subfigure depicts the low-dimensional projections of the NCCT volumes from the test set. As can be seen, a clustered structure emerged from the data points that was not related to the origin of the data or the presence of stroke lesions (colour of the dots).

Visual inspection was done on examples of each cluster, without identifying clear patterns, biases or shortcuts used by the model to aggregate the cases. In this sense, the middle and the bottom sub-figures show the middle slice of the 6 nearest neighbouring volumes from some representative points in the scatter plot (black markers). In the middle one, it can be seen that in general there are some common elements along the rows (neighbours). For example, in the first and second rows there is a similarity in the orientation of the volumes, and in the third row the cases seem to have large or abnormal ventricle sizes. The last sub-picture shows the attribution maps obtained from these cases. In the cases from the first row, the model was strongly influenced by some voxels in the frontal region, in the second row, some attention was given to the anterior voxels outside the brain (possibly related to the overall orientation of the case), and in the third row, the ventricular regions are highlighted. Overall, it is important to note that even when attention was paid to voxels outside the brain, the representations were able to capture information that was mostly related to the brain region.

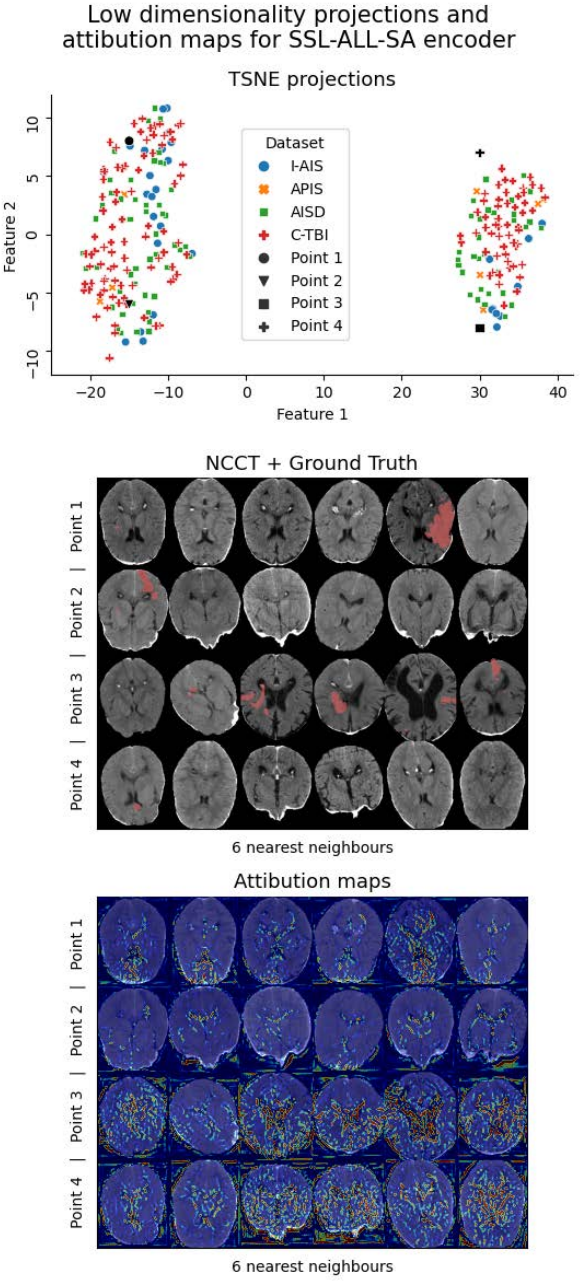


Figure 8: T-SNE low dimensional projections of the image representations and NCCT with the respective Attribution map for representative points.

5. Discussion

In this work, the use of a DeSD-like self-supervised pre-training strategy was proposed to exploit large unlabelled NCCT (stroke positive and negative) datasets in the task of AIS final infarct segmentation.

As a first remark, it is worth highlighting the software engineering contributions made in order to enhance the nnU-Net model with SSL pre-training of its encoder. In this work, an SSL training infrastructure was designed in such a way that it was automatically adapted to the guidelines resulting from the robust nnU-Net self-configuration pipeline. In this sense, the proposed method, represents a contribution beyond the problem or datasets addressed in this work, allowing the use of SSL pre-training in any other 3D medical image segmentation problem.

In the experiments conducted for AIS, the segmentation performance obtained by pre-training the nnU-Net encoder in a self-supervised learning fashion and then doing supervised segmentation training was significantly better than training nnU-Net from scratch on the same supervised learning dataset. These results were consistent with the findings of Ye et al. (2022), despite the different nature of the problem and the different network architecture used.

In terms of the dataset used to pre-train the encoder, it was interesting that only a small difference to was found between using all the stroke positive and negative cases and using only the AIS positive datasets, while both had a large performance gain with compared to using only the stroke negative cases for pre-training. There might be several explanations for this. One of them is that

the inclusion of stroke positive cases may be a key element in achieving a good pre-training for the supervised task at hand. However, the attention maps obtained for the encoder and the clusters shown by the low representations did not suggest that the pre-trained encoder itself captured any evident information about the stroke lesions. A more resonable explanation lies in the inclusion of cases in the training set for SSL that were also present in the supervised training set, allowing the model to take into account the particular image characteristics of this dataset. In any case, it was clear that the model did not benefit from the contrast of pre-training on healthy cases and then using only pathological ones for the supervised task.

In relation to the exploitation of contextual information, it was shown that including the proposed symmetry augmentation technique during pre-training led to a small improvement in the segmentation performance. This was not surprising, as it was consistent with many publications in the field showing that models achieve better AIS segmentation performance when forced to exploit inter-hemispheric symmetry/asymmetry. However, despite the initial belief that the symmetry augmentation would be more beneficial for the downstream task when pre-training on non-stroke cases (becoming normal asymmetry agnostic), the results obtained were exactly the opposite. As there is no immediate rational explanation for this phenomenon, further experiments should be conducted to better understand what is the effect of this augmentation on the obtained representations.

Also in relation to the contextual information, in the introduction it was initially hypothesised that the use of SSL could lead to meaningful representations of the NCCT images that could exploit the contextual information already present in the image itself. However, although the encoder's attribution maps showed that the pre-trained encoder was mostly influenced by brain voxels and the shape/location of certain structures such as the ventricles, it is difficult to determine the impact of this information on the final segmentation. The inclusion of additional pretext tasks during the SSL training, as done in Giancardo et al. (2023), is a strategy that could lead to representations better suited to the downstream problem and should be explored in future work.

One of the other important results of this work was that the best SSL pre-training strategy was able to achieve performances at least on par with the ones obtained by training the supervised segmentation model trained from scratch with almost 3.6 times more labelled cases. As commented previously, there was a high variability in the quality of the datasets, in their initial preprocessing, in the volume of the lesions in them, and only icoAIS had certain infarct ground truth masks. In this context, SSL pre-training provided a robust way of extracting meaningful information from these cases, becoming independent of their variable labelling standards.

5.1. Comparison with other approaches

From the results section, it is clear that models developed for other medical image segmentation problems achieve better Dice scores. However, AIS segmentation on NCCT is a particularly challenging task, due to the cross-domain nature of the labels, the lack of visibility of the lesions in NCCT, and many other reasons previously exposed.

Especially in the context of AIS, it is difficult to even compare with other approaches presented in the literature. In this work, the icoAIS dataset was chosen as the main dataset; this choice had a major drawback, which is the impossibility -in the time of the project- to compare our results with those of other methods, since neither their approaches nor our dataset are publicly available. However, this dataset was chosen because it was better than the publicly available ones in terms of size, type and quality of the labels. Therefore, in this work, the performance evaluation of the proposed methods was done by comparing the relative improvements of the proposed methods with a widely accepted baseline such as nnU-Net.

Having said this, and taking into account that the strict comparison with other reported methods may be misleading due to several differences (number of cases, label origin, minimum lesion size, lesion location, etc.), our results are in the same range as the those reported by other methods on datasets similar to ours. For example, in Kuang et al. (2021) a 3D UNet applied to AIS segmentation is reported to achieve a mean Dice of 0.308 (sd: 0.283), in Giancardo et al. (2023) their model achieved a mean Dice of 0.26 and a plain nnU-Net one of 0.14 and in El-Hariri et al. (2022) a 3D nnU-Net trained for AIS had a mean Dice of 0.377 or 0.346 (sd: 0.276 and 0.275 respectively) depending on the reader used as ground truth.

5.2. Limitations

Choosing which metric to report and focus on is not an easy task in AIS segmentation. Although the performance based on 95% Hausdorff distance and absolute volume difference was reported and commented in the results section, the focus in this work was placed on the Dice Score coefficient.

The 95% Hausdorff distance is not considered a clinically relevant metric in the field of AIS segmentation and was only included for consistency with other image analysis works. Due to the lack of contrast of AIS in NCCT images, it is hard to achieve accurate contour matching, making it very difficult to get a clear picture of the overall model performance using HD95. AVD, on the other hand, is a very relevant measure from a clinical perspective, nevertheless the values obtained for it should be put into context. With Dice values not surpassing 0.35, it is difficult to tell if a model is better than

another simply because it achieved a smaller volume difference. A smaller AVD indicates that the volumes are similar, but in our problem they are most likely mislocalised, so this should be interpreted roughly linked to how specific the models are.

Finally, DSC is not without its complications. The Dice score is biased by the size of the lesion volume, i.e. low spatial overlap for a big lesion might generate a high Dice and vice versa. As a consequence, the increase in DSC reported in the results for all models when the lesions smaller than 3mL were removed from the ground truth, could have two explanations. On the one hand, it could indicate that all models struggled to segment very small lesions, but on the other hand, it could be a consequence of the limitations of the metric itself. Due to the wide range of lesion volumes, depending on the case, small lesions may have a large influence on the metric, which is inconsistent with the lower clinical significance assigned to them.

Although Dice was chosen as the metric to focus on, all of the above concerns should be taken into account when interpreting the results obtained. For future work, a ranking system, such as the one implemented in (Maier et al. (2017)), could be used to collect and ponder the information provided by the different metrics.

Regarding the experimental design, some pitfalls in the dataset partitioning strategy need to be pointed out. Firstly, the distribution of cases between the training, validation and test sets could have been done better. A minimum number of cases should have been guaranteed to be left in the test set to allow for more powerful statistical analysis of the results.

Secondly, as previously commented, in this work a stratified partitioning was done according to data origin, lesion location and lesion size. However, it might have been advantageous not to restrict the location to the sides of the brain, but to specify the nervous system structures affected, as lesions in the cerebellum and brain stem may be more difficult to segment due to bone-related imaging artefacts.

Lastly, more robust results could have been obtained by validating the models using a k-fold cross-validation procedure. However, in a context of limited computational resources, it was preferred to run each supervised nnU-Net experiment multiple times and enssembling the resulting models to reduce the impact of nnU-Net random initialisation on segmentation performance.

6. Conclusions

In this work, a DeSD-like self-supervised pre-training strategy was proposed to exploit large unlabelled NCCT (stroke positive and negative) datasets in the task of AIS final infarct segmentation. A robust data pre-processing pipeline was proposed to homogenise the different datasets before using them in an SSL-enhanced version of the well-known self-configuring nnU-Net model.

From the conducted experiments, pre-training the nnU-Net's encoder in a self-supervised manner with all the available NCCT images (stroke-positive and stroke-negative) resulted in an AIS segmentation performance significantly higher than training the same model from scratch and comparable to that obtained by using approximately 3.6 times more labelled data.

In acute ischemic stroke, is very difficult to have access access to high quality datasets with both baseline and follow up images to accurately asses the extension of the final infarct (or ischemic lesion core). In this work, we presented a successful method to exploit large amounts of unlabelled baseline NCCT images, which are much easier to obtain from hospitals and are currently neglected, proving they can be used to improve final infarct lesion segmentation.

Acknowledgements

I would like to express my deepest gratitude to my supervisors for their trust, feedback and for the freedom they gave me throughout this project. I am also grateful to my family and friends who, despite the physical distances, have always been my emotional support. Lastly, I would like to mention Enzo Ferrante, a crucial link in a long chain of events that made this thesis possible.

Finally, this endeavour would not have been possible without the financial support of the Erasmus Mundus programme of the European Union and the computing resources provided by *icometrix*, Leuven, Belgium.

References

Albers, G.W., Marks, M.P., Kemp, S., Christensen, S., Tsai, J.P., Ortega-Gutierrez, S., McTaggart, R.A., Torbey, M.T., Kim-Tenser, M., Leslie-Mazwi, T., et al., 2018. Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging. New England Journal of Medicine 378, 708–718. doi:10.1056/nejmoa1713973.

Amador, K., Wilms, M., Winder, A., Fiehler, J., Forkert, N., 2021. Stroke lesion outcome prediction based on 4d CT perfusion data using temporal convolutional networks, in: Proceedings of the Fourth Conference on Medical Imaging with Deep Learning, PMLR. pp. 22–33.

Amador, K., Winder, A., Fiehler, J., Wilms, M., Forkert, N.D., 2022. Hybrid spatio-temporal transformer network for predicting ischemic stroke lesion outcomes from 4d CT perfusion imaging. Lecture Notes in Computer Science , 644–654doi:10.1007/978-3-031-16437-8_62.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. NeuroImage 26, 839–851. doi:10.1016/j.neuroimage.2005.02.018.

Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y., Schwarzschild, A., Wilson, A., Geiping, J., Garrido, Q., Fernandez, P., Bar, A., Pirsiavash, H., Lecun, Y., Goldblum, M., 2023. A cookbook of self-supervised learning .

Bardes, A., Ponce, J., LeCun, Y., 2022. VICReg: Variance-Invariance-Covariance Regularization for self-supervised learning, in: International Conference on Learning Representations. URL: .

Benjamin, E.J., Blaha, M.J., Chiuve, S.E., Cushman, M., Das, S.R., Deo, R., de Ferranti, S.D., Floyd, J., Fornage, M., Gillespie, C., et al., 2017. Heart disease and stroke statistics—2017 update: A report from the American Heart Association. Circulation 135. doi:10.1161/cir.0000000000000485.

Bouslama, M., Ravindran, K., Harston, G., Rodrigues, G.M., Pisani, L., Haussen, D.C., Frankel, M.R., Nogueira, R.G., 2021. Noncontrast computed tomography e-stroke infarct volume is similar to rapid computed tomography perfusion in estimating postreperfusion infarct volumes. Stroke 52, 634–641. doi:10.1161/strokeaha.120.031651.

Brorson, J.R., Cifu, A.S., 2019. Management of Patients With Acute Ischemic Stroke. JAMA 322, 777–778. doi:10.1001/jama.2019.10436.

Byrne, D., Walsh, J.P., MacMahon, P.J., 2019. An acute stroke CT imaging algorithm incorporating automated perfusion analysis. Emergency Radiology 26, 319–329. doi:10.1007/s10140-019-01675-2.

Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised Vision Transformers. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) doi:10.1109/iccv48922.2021.00951.

Chen, W., Wu, J., Wei, R., Wu, S., Xia, C., Wang, D., Liu, D., Zheng, L., Zou, T., Li, R., et al., 2022. Improving the diagnosis of acute ischemic stroke on non-contrast CT using Deep Learning: A multicenter study. Insights into Imaging 13. doi:10.1186/s13244-022-01331-3.

Chen, X., He, K., 2021. Exploring simple siamese representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15750–15758.

Clèrigues, A., Valverde, S., Bernal, J., Freixenet, J., Oliver, A., Lladó, X., 2020. Acute and sub-acute stroke lesion segmentation from multimodal MRI. Computer Methods and Programs in Biomedicine 194, 105521. doi:10.1016/j.cmpb.2020.105521.

El-Hariri, H., Souto Maior Neto, L.A., Cimflova, P., Bala, F., Golan, R., Sojoudi, A., Duszynski, C., Elebute, I., Mousavi, S.H., Qiu, W., et al., 2022. Evaluating nnU-net for early ischemic change segmentation on non-contrast computed tomography in patients with acute ischemic stroke. Computers in Biology and Medicine 141, 105033. doi:10.1016/j.compbiomed.2021.105033.

Estrada, U.M., Meeks, G., Salazar-Marioni, S., Scalzo, F., Farooqui, M., Vivanco-Suarez, J., Gutierrez, S.O., Sheth, S.A., Giancardo, L., 2022. Quantification of infarct core signal using CT imaging in acute ischemic stroke. NeuroImage: Clinical 34, 102998. doi:10.1016/j.nicl.2022.102998.

Farzin, B., Fahed, R., Guilbert, F., Poppe, A.Y., Daneault, N., Durocher, A.P., Lanthier, S., Boudjani, H., Khoury, N.N., Roy, D., et al., 2016. Early CT changes in patients admitted for thrombectomy. Neurology 87, 249–256. doi:10.1212/wnl.0000000000002860.

Garrido, Q., Balestriero, R., Najman, L., Lecun, Y., 2023. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank doi:10.48550/arXiv.2210.02885.

Giancardo, L., Niktabe, A., Ocasio, L., Abdelkhaleq, R., Salazar-Marioni, S., Sheth, S.A., 2023. Segmentation of acute stroke infarct core using image-level labels on CT-Angiography. NeuroImage: Clinical 37, 103362. doi:10.1016/j.nicl.2023.103362.

Goyal, M., Ospel, J.M., Menon, B., Almekhlafi, M., Jayaraman, M., Fiehler, J., Psychogios, M., Chapot, R., van der Lugt, A., Liu, J., et al., 2020. Challenging the ischemic core concept in acute ischemic stroke imaging. Stroke 51, 3147–3155. doi:10.1161/strokeaha.120.030620.

Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., 2020. Bootstrap your own latent a new approach to self-supervised learning, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA.

Gómez, S., Florez, S., Mantilla, D., Camacho, P., Tarazona, N., Martínez, F., 2023. An attentional U-Net with an auxiliary class learning to support acute ischemic stroke segmentation on CT. Medical Imaging 2023: Image Processing doi:10.1117/12.2654269.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16000–16009.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Hernandez Petzsche, M.R., de la Rosa, E., Hanning, U., Wiest, R., Valenzuela, W., Reyes, M., Meyer, M., Liew, S.L., Kofler, F., Ezhov, I., et al., 2022. Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. Scientific Data 9. doi:10.1038/s41597-022-01875-5.

Hoopes, A., Mora, J.S., Dalca, A.V., Fischl, B., Hoffmann, M., 2022. Synthstrip: Skull-stripping for any brain image. NeuroImage 260, 119474. doi:10.1016/j.neuroimage.2022.119474.

Iglesias, J.E., Billot, B., Balbastre, Y., Magdamo, C., Arnold, S.E., Das, S., Edlow, B.L., Alexander, D.C., Golland, P., Fischl, B., 2023. SynthSR: A public AI tool to turn heterogeneous clinical brain scans into high-resolution T1-weighted images for 3D morphometry. Science Advances 9. doi:10.1126/sciadv.add3607.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2020. nnU-net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18, 203–211. doi:10.1038/s41592-020-01008-z.

Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H., Heiland, S., Wick, W., et al., 2019. Automated brain extraction of multisequence MRI using artificial neural networks. Human Brain Mapping 40, 4952–4964. doi:10.1002/hbm.24750.

Jiang, Y., Sun, M., Guo, H., Yan, K., Lu, L., Xu, M., 2023. Anatomical invariance modeling and semantic alignment for self-supervised learning in 3D medical image segmentation. arXiv preprint arXiv:2302.05615 .

Kalapos, A., Gyires-Tóth, B., 2023. Self-supervised pretraining for 2D medical image segmentation. Lecture Notes in Computer Science , 472–484doi:10.1007/978-3-031-25082-8_31.

Kim, Y., Lee, S., Abdelkhaleq, R., Lopez-Rivera, V., Navi, B., Kamel, H., Savitz, S.I., Czap, A.L., Grotta, J.C., McCullough, L.D., et al., 2021. Utilization and availability of advanced imaging in patients with acute ischemic stroke. Circulation: Cardiovascular Quality and Outcomes 14. doi:10.1161/circoutcomes.120.006989.

Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine 15, 155–163. doi:10.1016/j.jcm.2016.02.012.

Kuang, H., Menon, B.K., Qiu, W., 2019. Automated infarct segmentation from follow-up non-contrast CT scans in patients with acute ischemic stroke using dense multi-path contextual generative adversarial network. Lecture Notes in Computer Science , 856–863doi:10.1007/978-3-030-32248-9_95.

Kuang, H., Menon, B.K., Sohn, S.I., Qiu, W., 2021. EIS-net: Segmenting early infarct and scoring aspects simultaneously on non-contrast CT of patients with acute ischemic stroke. Medical Image Analysis 70, 101984. doi:10.1016/j.media.2021.101984.

Li, S., Zheng, J., Li, D., 2021. Precise segmentation of non-enhanced computed tomography in patients with ischemic stroke based on multi-scale U-net Deep Network model. Computer Methods and Programs in Biomedicine 208, 106278. doi:10.1016/j.cmpb.2021.106278.

Li, X., Morgan, P.S., Ashburner, J., Smith, J., Rorden, C., 2016. The first step for neuroimaging data analysis: DICOM to NIFTI conversion. Journal of Neuroscience Methods 264, 47–56. doi:10.1016/j.jneumeth.2016.03.001.

Lutkenhoff, E.S., Rosenberg, M., Chiang, J., Zhang, K., Pickard, J.D., Owen, A.M., Monti, M.M., 2014. Optimized brain extraction for pathological brains (optibet). PLoS ONE 9. doi:10.1371/journal.pone.0115551.

Maas, A.I., Menon, D.K., Steyerberg, E.W., Citerio, G., Lecky, F., Manley, G.T., Hill, S., Legrand, V., Sorgner, A., 2015. Collaborative european neurotrauma effectiveness research in traumatic brain injury (center-TBI). Neurosurgery 76, 67–80. doi:10.1227/neu.0000000000000575.

van der Maaten, L., Hinton, G., 2008. Viualizing data using t-SNE. Journal of Machine Learning Research 9, 2579–2605.

Maier, O., Menze, B.H., von der Gablentz, J., Häni, L., Heinrich, M.P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., et al., 2017. ISLES 2015 - a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. Medical Image Analysis 35, 250–269. doi:10.1016/j.media.2016.07.009.

Manna, S., Chakraborty, S., 2022. BYOLMed3D: Self-supervised representation learning of medical videos using gradient accumulation assisted 3D BYOL framework doi:10.48550/arXiv.2208.00444.

Mokin, M., Primiani, C.T., Siddiqui, A.H., Turk, A.S., 2017. ASPECTS (Alberta stroke program early CT score) measurement using Hounsfield unit values when selecting patients for stroke thrombectomy. Stroke 48, 1574–1579. doi:10.1161/strokeaha.117.016745.

Ni, H., Xue, Y., Wong, K., Volpi, J., Wong, S.T., Wang, J.Z., Huang, X., 2022. Asymmetry disentanglement network for interpretable acute ischemic stroke infarct segmentation in non-contrast CT scans. Lecture Notes in Computer Science , 416–426doi:10.1007/978-3-031-16452-1_40.

Nogueira, R.G., Jadhav, A.P., Haussen, D.C., Bonafe, A., Budzik, R.F., Bhuva, P., Yavagal, D.R., Ribo, M., Cognard, C., Hanel, R.A., et al., 2018. Thrombectomy 6 to 24 hours after stroke with a mismatch between deficit and infarct. New England Journal of Medicine 378, 11–21. doi:10.1056/nejmoa1706442.

Ostmeier, S., Heit, J., Axelrod, B., Li, L.J., Zaharchuk, G., Verhaaren, B., Mahammedi, A., Christensen, S., Lansberg, M., 2022. Non-inferiority of deep learning model to segment acute stroke on non-contrast CT compared to neuroradiologists doi:10.48550/arXiv.2211.15341.

Powers, W.J., Rabinstein, A.A., Ackerson, T., Adeoye, O.M., Bambakidis, N.C., Becker, K., Biller, J., Brown, M., Demaerschalk, B.M., Hoh, B., et al., 2019. Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke. Stroke 50. doi:10.1161/str.0000000000000211.

Robben, D., Boers, A.M., Marquering, H.A., Langezaal, L.L., Roos, Y.B., van Oostenbrugge, R.J., van Zwam, W.H., Dippel, D.W., Majoie, C.B., van der Lugt, A., et al., 2020. Prediction of final infarct volume from native CT perfusion and treatment parameters using deep learning. Medical Image Analysis 59, 101589. doi:10.1016/j.media.2019.101589.

Rorden, C., Bonilha, L., Fridriksson, J., Bender, B., Karnath, H.O., 2012. Age-specific CT and MRI templates for spatial normalization. NeuroImage 61, 957–965. doi:10.1016/j.neuroimage.2012.03.020.

Sacco, R.L., Kasner, S.E., Broderick, J.P., Caplan, L.R., Connors, J.B., Culebras, A., Elkind, M.S., George, M.G., Hamdan, A.D., Higashida, R.T., et al., 2013. An updated definition of stroke for the 21st Century. Stroke 44, 2064–2089. doi:10.1161/str.0b013e318296aeca.

Shamonin, D., 2013. Fast parallel image registration on CPU and GPU for diagnostic classification of alzheimer's disease. Frontiers in Neuroinformatics 7. doi:10.3389/fninf.2013.00050.

Sotoudeh, H., Bag, A.K., Brooks, M.D., 2019. "code-stroke" CT perfusion; challenges and pitfalls. Academic Radiology 26, 1565–1579. doi:10.1016/j.acra.2018.12.013.

Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic attribution for Deep Networks, in: Precup, D., Teh, Y.W. (Eds.), Proceedings of the 34th International Conference on Machine Learning, PMLR. pp. 3319–3328.

Vagal, A., Wintermark, M., Nael, K., Bivard, A., Parsons, M., Grossman, A.W., Khatri, P., 2019. Automated CT perfusion imaging for acute ischemic stroke. Neurology 93, 888–898. doi:10.1212/wnl.0000000000008481.

Virani, S.S., Alonso, A., Aparicio, H.J., Benjamin, E.J., Bittencourt, M.S., Callaway, C.W., Carson, A.P., Chamberlain, A.M., Cheng, S., Delling, F.N., et al., 2021. Heart disease and stroke statistics—2021 update. Circulation 143. doi:10.1161/cir.0000000000000950.

Ye, Y., Zhang, J., Chen, Z., Xia, Y., 2022. DeSD: Self-supervised learning with deep self-distillation for 3D medical image segmentation. Lecture Notes in Computer Science , 545–555doi:10.1007/978-3-031-16440-8$_5$2.

Appendix A. ISBI 2023 APIS Challenge solution

As mentioned in Section 3.7.4, our submission to the ISBI 2023 APIS Challenge won the competition. The proposed model achieved a Dice score of 0.20 ± 0.25 and a Hausdorff distance of 66.02 ± 24.22 on the hidden test set, which were significantly better than those of the second best solution (0.11 ± 0.30 and 59.64 ± 22.88, respectively). Since both solutions used a 3D nnU-Net model, the difference in performance was mainly explained by the use of the robust preprocessing strategy presented in Section 3.2, which allowed us to use a larger set of training cases, and by the inclusion of the difference image as an additional input channel.

To obtain this model, the inter-hemispheric difference image was first generated by subtracting the NCCT from its contralateral version, resulting in an image in which both normal and abnormal inter-hemispheric differences appeared highlighted. Then, a training procedure was performed in the same way as that described in Section 3.7.1 for the baseline *FS-STKp*.

Appendix B. Addition of interhemispherical difference image as another input channel

In line with Appendix A, for this experiment the interhemispheric difference image was generated and then used as an additional input channel in both the SSL pretraining and the supervised experiments.

The two baselines presented in Section 3.7.1 were run for the multi-channel input. Additionally, the nnU-Net encoder was SSL pre-trained with the three dataset configurations presented in Section 3.7.2. Once the encoder was pre-trained, the full supervised nnU-Net was trained using only the icoAIS dataset. All training procedures followed exactly the same specifications as described in section 3.

Tables A.4 and A.5 show the quantitative results obtained for the segmentation task on the icoAIS validation set. Contrary to the results presented in Section 4, it can be seen here that pre-training the nnU-Net encoder with SSL and fine-tuning the full architecture with supervised training led to worse performance than training nnU-Net from scratch on the same dataset.

Comparing the results between the supervised models trained from scratch with and without the inclusion of the difference image, it can be seen that, in line with the

Table B.4: Performance measures for multichannel input with difference image + NCCT, on (icoAIS val. set) considering all lesions sizes (n=29).

	DSC ↑		HD95 ↓		AVD ↓		Corr ↑	ICC ↑
Experiment	Mean	Median(Iqr)	Mean	Median(Iqr)	Mean	Median(Iqr)		
FS-STKi	**0.2445**	0.130 (0.470)	44.39	41.15 (32.84)	**22.00**	15.77 (23.23)	**0.7**	**0.69**
FS-STKp	0.2380	**0.185** (0.420)	**38.52**	**34.05** (19.33)	24.48	**14.04** (23.74)	0.69	0.60
ALL-STKi	0.1897	0.122 (0.267)	57.20	55.66 (35.55)	51.22	43.82 (33.68)	0.46	0.26
STKp-STKi	0.1989	0.139 (0.303)	53.51	54.48 (21.97)	47.01	34.42 (33.55)	0.34	0.30
STKn-STKi	0.1953	0.119 (0.332)	55.91	54.08 (26.03)	58.09	48.42 (50.82)	0.42	0.25

Table B.5: Performance measures for multichannel input with difference image + NCCT, on *icoAIS val. set* considering lesions > 3*mL* (n=24).

	DSC ↑		HD95 ↓		AVD ↓		Corr ↑	ICC ↑
Experiment	Mean	Median(Iqr)	Mean	Median(Iqr)	Mean	Median(Iqr)		
FS-STKi	0.2890	**0.314** (0.435)	41.78	**36.47** (34.01)	**25.04**	**12.47** (25.68)	0.57	**0.64**
FS-STKp	**0.2920**	0.293 (0.495)	**41.63**	36.60 (38.42)	25.21	17.76 (27.50)	**0.74**	0.58
ALL-STKi	0.2253	0.144 (0.238)	64.73	66.40 (37.36)	53.25	33.67 (35.06)	0.28	0.17
STKp-STKi	0.2361	0.162 (0.285)	59.81	62.44 (27.27)	48.25	33.36 (39.39)	0.23	0.21
STKn-STKi	0.2317	0.149 (0.295)	62.27	62.78 (32.34)	60.70	42.83 (58.37)	0.26	0.16

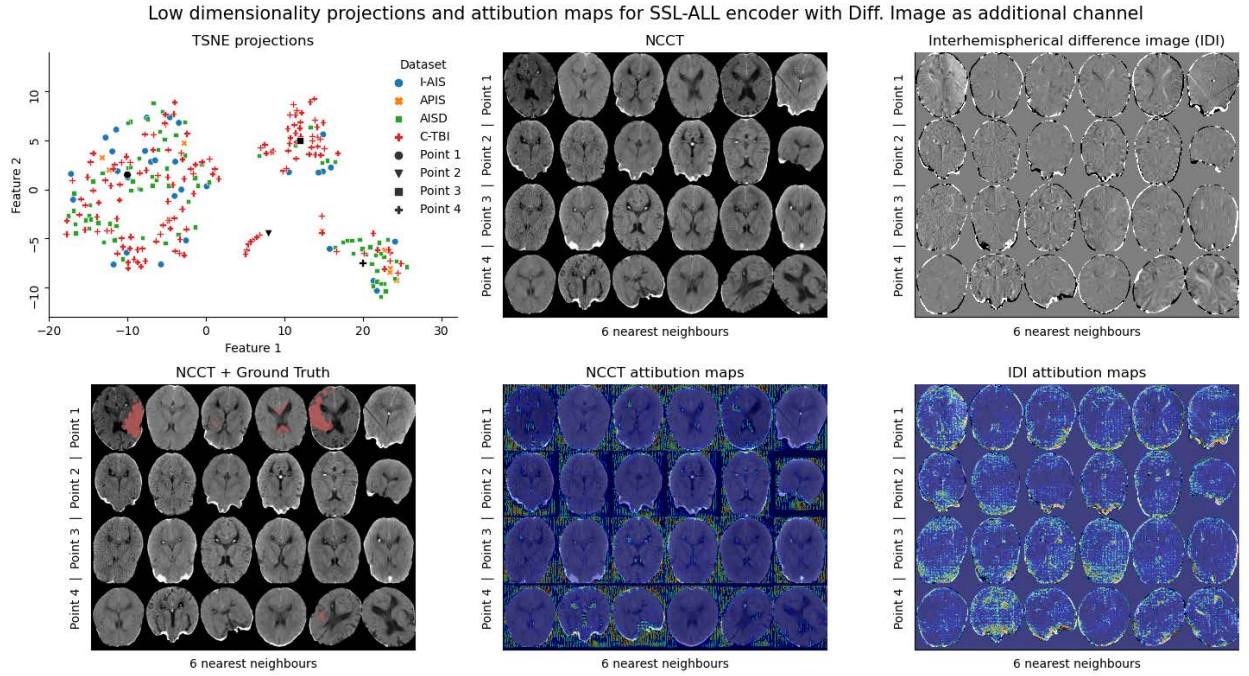


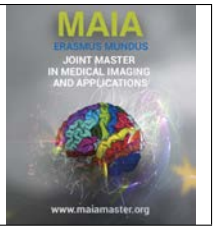
Figure B.9: T-SNE low dimensional projections of the image representations and NCCT and interhemispherical difference images examples (with their the respective attribution maps) for representative points.

literature, the inclusion of this additional input channel was beneficial.

As can be seen in Figure B.9, the attention maps of the encoders pre-trained with SSL show that the output of the model was strongly influenced by the brain region in the difference image channel and, unexpectedly, by the non-brain region of the NCCT channel. This finding, in addition to the benefits seen in training from scratch, suggests that further experiments should be conducted to explore different ways of exploiting the potential of the two images during self-supervised pre-training (i.e. as data augmentation techniques, separate paths for each image, etc.).

Medical Imaging and Applications

Master Thesis, June 2023



Hemorrhagic stroke hematoma expansion detection and prediction using non-contrast computed tomography images

Cansu Yalçın[a], Valeriia Abramova[a], Yolanda Silva[b], Arnau Oliver[a], Xavier Lladó[a]

[a] *Computer Vision and Robotics Group, University of Girona, Spain*
[b] *Stroke Unit, Department of Neurology, Hospital Universitari Doctor Josep Trueta de Girona, Girona, Spain*

Abstract

Hemorrhagic stroke refers to bleeding when a blood vessel in the brain ruptures, leading to the formation of a hematoma as the blood flows into the surrounding brain tissue. Due to its high mortality rates, a quick response is crucial to prevent irreversible consequences. Hematoma expansion (HE) is a term to describe the rise in the volume of the hematoma over time. HE is characterized by either a rise in absolute volume of more than 6 ml or a relative volume increase of over 33% on the follow-up non-contrast computed tomography (NCCT) scan when compared to the initial scan. Presence of the hematoma growth is related to the worsening of the clinical outcomes. Therefore, accurately identifying patients at risk is critical, since they could be targeted for clinical treatment. In this dissertation, we propose two deep learning-based applications solely using NCCT data. First, a novel hematoma growth detection approach to automatically measure the growth from the longitudinal CT scans of a patient, having two distinct time points (basal and follow-up). Second, a prediction for the HE occurrence using only information from the basal image. We have studied various deep learning models such as modified Unet encoders with attention gate (Unet-AG) and squeeze and excitation blocks (Unet-SE), transfer learning models (Densenet, EfficientNet, Resnet), and vision transformers (Swin-t and R50-ViT). In our study, we conducted experiments using both 2D and 3D settings, aiming to assess the advantages and disadvantages of these approaches within the same dataset. Furthermore, we experimented with different input variations such as whole-image, ROI-based and lesion-based approaches. All analyses have been performed using a five-fold cross-validation strategy using the dataset obtained from Hospital Dr. Josep Trueta, which consisted of 206 cases, out of which 41 were confirmed HE cases and the rest were negative cases, i.e. cases without or with small hematoma expansion. The overall performance results of the detection models were as follows: the 2D detection model utilizing the Unet approach achieved a ROC-AUC score of 0.920 with both ROI-based image inputs. The 3D detection approach using the Unet model and ROI-based image input achieved a ROC-AUC score of 0.800. Lastly, for the challenging 2D prediction model using only basal images the use of the EfficientNetB0 model and whole image input achieved a ROC-AUC score of 0.720. The obtained results show promising potential to be explored in the clinical setting.

Keywords: Hemorrhagic stroke, Hematoma expansion, Lesion growth, Classification, Deep learning

1. Introduction

Nowadays stroke is one of the most common causes of death, ranking fifth place among all causes of death after diseases of the heart, cancer, COVID-19, and unintentional injuries (Tsao., 2023). Stroke is a medical condition that occurs when the blood supply to a part of the brain is disrupted or reduced. This can be caused by a blockage in a blood vessel or bleeding in the brain. When the brain is deprived of oxygen and nutrients, brain cells begin to die within minutes, which can lead to permanent brain damage or even death. Stroke is classified into two types: ischemic and intracerebral hemorrhage (ICH). Ischemic stroke occurs when a blood clot blocks a blood vessel in the brain, cutting off the blood supply to that area of the brain. ICH occurs when a blood vessel in the brain ruptures or leaks, causing bleeding in the brain. Ischemic stroke is the

most common type of stroke, accounting for about 87% of all strokes and ICH is less common, for about 10% of all strokes (Tsao., 2023). Even though it is a less common condition, it has the highest death rates (Rost et al., 2008), with approximately 40% one-month mortality rates (Qureshi et al., 2009).

Within the occurrence of ICH, blood can flow into the surrounding brain tissue if a blood artery in the brain bursts, creating a hematoma. Hematoma expansion (HE) is the term used to describe a rise in the size of a hematoma, or collection of blood, inside the brain following a hemorrhagic stroke. The hematoma may continue to enlarge as more blood pours from the broken blood vessel over time. HE is characterized by either a rise in absolute volume of more than 6 ml or a relative volume increase of over 33% on the follow-up CT scan when compared to the initial scan (Wada et al., 2007). HE affects 20–40% of hemorrhagic stroke patients and it occurs within 24 hours of ICH (Davis et al., 2006), (Dowlatshahi et al., 2011). Previous studies showed that the presence of growth is related to the worsening of clinical outcomes and an increase in mortality (Steiner et al., 2006). As a result, accurately identifying patients at risk for early expansion is critical, since they could be a target for clinical treatment.

Neuroimaging is the main method in the diagnosis of ICH. Computed tomography (CT) imaging is fast, inexpensive, and widely available, these features support it to be the most common choice for the neuroimaging of stroke patients. CT angiography (CTA) is an imaging technique that can be used in the neuroimaging of ICH cases. CTA is used to visualize blood vessels in the body, particularly the arteries. It combines CT scanning with contrast material to produce detailed, three-dimensional images of blood vessels. However, for some patients with special conditions, the use of CTA may be limited due to a variety of factors. Therefore, non-contrast computed tomography (NCCT) is the most frequently used modality for detecting ICH, identifying prognostic indicators, and measuring hematoma volume, all of which are critical for the prognosis of the disease (Hillal et al., 2022). Notice that two images need to be acquired in the last case, the basal and the follow-up, usually acquired within 8 hours from the onset for basal and for follow-up 24 hours later after baseline.

Recently, the prediction of hematoma expansion using only the basal image has drawn increasing attention in research. NCCT imaging markers such as hypodensity, black hole sign, blend sign, island sign, and swirl sign has been proposed to be indicators of hematoma expansion (Cai et al., 2020). Examples of these imaging signs are given in Figure 1. While these signs can be helpful in identifying hemorrhages, they are not always accurate or present in the images. The sensitivity, or the ability of these signs to correctly identify a hematoma, is only about 50% (Tingting et al., 2022). This means that half of the time, these signs may not be present even if there is a hemorrhage, making it difficult to accurately identify the expansion of the hematoma.

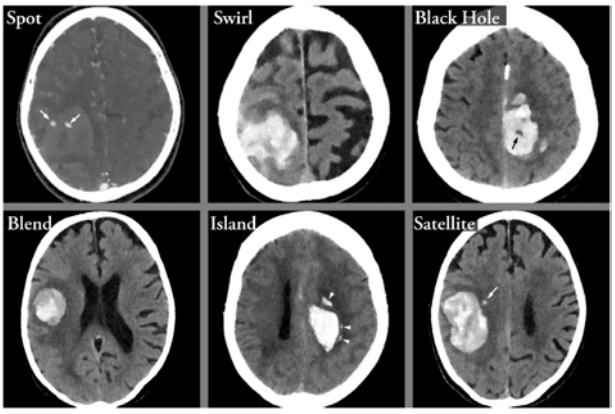


Figure 1: Imaging signs of hematoma expansion. Image extracted from Hillal et al. (2022).

In clinical settings, NCCT imaging markers are analyzed for the prediction of hematoma expansion by experienced radiologists. Interpreting these CT signals depends primarily on skilled radiologists, making it costly and labor-intensive. As a result, automated hematoma expansion prediction is critical. However, predicting hematoma expansion remains difficult due to the complex relationship between hematoma expansion and its various factors, along with the considerable variability between individual hematoma conditions. Nonetheless, artificial intelligence (AI) can offer a solution by analyzing vast amounts of medical data and identifying patterns that may be difficult for clinicians to detect. By leveraging the power of AI, medical professionals can more quickly and accurately diagnose hematoma expansion, allowing for earlier treatment and better patient outcomes.

1.1. Contributions

In this study, we have explored two approaches. First, a novel growth detection approach to automatically measure the growth from the longitudinal CT scans of a patient, having two distinct time points (basal and the follow-up). This application targets detecting the HE without the need for segmentation masks, leading to a rapid response. Second, a prediction approach for the HE using only basal image information. We studied different deep learning models in different image approaches (whole image, ROI-based and lesion-based) in 2D and 3D settings.

The following points provide a concise summary of the key contributions made during the development of our detection and prediction models:

1. We have studied various deep learning models such as modified Unet encoders with attention gate and squeeze-and-excitation blocks, transfer learning models (Densenet, EfficientNet, Resnet), and vision transformers (Swin, R50-ViT).
2. We explored the challenging problem of prediction of HE using only basal image as an input.
3. We have performed our experiments in 2D and 3D settings, allowing us to compare the advantages and disadvantages of small and imbalanced datasets.
4. We have experimented with different input variations of the data such as whole-image, ROI-based and lesion-based approaches. The implementations were made using an in-house dataset from the Trueta Hospital located in Girona, Spain.
5. We investigated the impact of excluding patients with intraventricular hemorrhage (IVH) in deep learning models.
6. In contrast to much of the existing literature, our algorithm solely utilizes non-contrast computed tomography (NCCT) image data. This has the potential to advance research toward image-only data utilization, which is often the primary data source in emergency situations.

2. State of the art

A multitude of investigations has been carried out on the application of machine learning and deep learning algorithms for the analysis of medical data to predict HE. Given the rarity of hemorrhagic stroke, gathering large amounts of data is difficult. To the best of our knowledge, there are no available public datasets specifically designed for hematoma expansion prediction. Consequently, existing research in the literature has relied on limited private datasets from local hospitals. This situation unfortunately introduces a bias towards a small patient population, making it challenging to directly compare results between different studies. In other words, while the methodology used may be consistent, the success of the models heavily relies on the characteristics of the dataset employed.

Our literature review shows that prior studies on hematoma expansion have primarily used machine learning classifiers that use imaging markers, radiomics features, and clinical data information. With recent advances in deep learning techniques, there has been an increase in interest in using deep neural networks (DNNs) to classify cases of hematoma expansion, resulting in a growth in the number of proposed algorithms.

2.1. Machine learning based analyses in HE prediction

In light of the success of conventional machine learning algorithms on small datasets, it is observed that machine learning has been widely utilized, with recent research continuing to demonstrate its applications in the prediction of hematoma expansion. For instance, Liu et al. (2019) implemented a support vector machine classification application by combining different variables, including the patient's demographic parameters, clinical status, laboratory test parameters, and image signs. The work of Duan et al. (2021) compared radiomic models based on different machine learning models for the hematoma expansion prediction such as support vector machine (SVM), decision tree (DT), conditional inference trees (CIT), random forest (RF), k-nearest neighbors (KNN), back-propagation neural network (BPNet) and Bayes. They used texture parameters from the baseline NCCT images as inputs. In their study, they excluded intraventricular hemorrhage cases. In their recent work, Chen et al. (2022) conducted a comparative analysis of three models aimed at predicting hematoma expansion. The models incorporated radiomics features extracted from the NCCT image, radiological features manually defined by two experienced radiologists, and a combined model leveraging both types of inputs. The authors evaluated the performance of the Catboost model in comparison with other traditional machine learning models, as utilized in previous studies. As another recent traditional machine learning application, Li et al. (2023) extracted radiomics features that were categorized into three groups based on geometry, intensity, and texture. The authors conducted Lasso feature screening and prediction analysis using eight distinct machine learning models.

2.2. Deep learning based analyses in HE prediction

There has been a surge of interest in the use of deep learning for predicting hematoma expansion. It is worth noting that, in the literature, deep learning techniques have primarily been employed on modestly sized architectures with fewer parameters, due to the size of the available datasets. In addition, we have observed a trend towards incorporating other forms of information, such as clinical data or image features that are derived through traditional image processing techniques, in conjunction with medical imaging data.

Several studies in the literature have integrated medical imaging data with clinical parameters. For instance, Wang et al. (2021) presented an automated prediction pipeline utilizing NCCT images and clinical parameters. Their architecture features a channel-attention-based encoder to extract image features and a decoder branch to upsample the clinical parameter features, followed by a fusion of the two sets of information. As another example of usage of clinical parameters, Wan et al. (2022) proposed BSGNet, which uses multimodal data to predict hematoma expansion and has a lower computational complexity. In their approach, the authors

achieved joint training with imaging data and clinical parameters.

In the literature, it was also possible to observe examples of the combination of machine learning and deep learning methods. For instance, Tingting et al. (2022) proposed a dual model machine learning method. The first step is to create a deep neural network predictor using Resnet-34, VGGNet, and GoogLeNet, using cropped lesion ROIs as input. In the second step, they combined the predictions with the clinical data of the patients and then had an MLP layer for the final classification.

There are also examples of using only image-based data in the existing literature. For instance, Zhong et al. (2021) presented a 3D Unet-like convolutional neural network (CNN) model for predicting hematoma expansion in their recent study. The authors compared the efficacy of conventional NCCT markers, a BAT predictive model which is a multi-itemed score for HE prediction (B for blend sign, A for hypodensity presence, and T for the time from onset to NCCT), and a deep learning model, concluding that the deep learning model outperformed the others.

Ma et al. (2022) proposed an end-to-end deep learning algorithm for the segmentation of hematoma lesions and prediction of hematoma expansion in 2D. They calculated evaluation metrics patient-wise, by taking the mean value of the given slice probabilities belonging to the same patient. The authors reported that the 2D Unet model with attention for the segmentation task and the 2D Resnet-34 model for the classification task yielded the best results.In another study, Tang et al. (2022) presented a model comprising the k-nearest neighbors matting method for skull stripping and a modified 2D Resnet-34 model for classification. They calculated evaluation metrics image-wise. The prediction model proposed by Teng et al. (2021) involved a combination of radiomic features extracted from the images and CNN features obtained using an Unet-like model in 2D. The combined features were fed into a Gradient Boosting classifier, which assigned the predictions.

Within the literature, there have been successful applications for predicting hematoma expansion utilizing solely NCCT images in both 2D (Ma et al. (2022), Teng et al. (2021), Tang et al. (2022)) and 3D (Zhong et al. (2021)) scenarios. However, there exists a dearth of research that comprehensively analyzes and compares these 2D and 3D approaches on the same dataset. Additionally, we observe a mixture of image-wise and patient-wise metric calculations in 2D applications, with no examples of comparison between these two disparate metrics in the current literature.

Conversely, while there are instances in the literature that employ cropped lesion regions of interest (ROIs) (Tingting et al. (2022)) and entire volume/image data, there is a lack of analysis and understanding of how different scales of NCCT image information impact the performance of deep learning-based models in predicting hematoma expansion.

2.3. Lesion detection

Longitudinal lesion detection is an analysis of detecting the worsening of the lesion at given two-time points. In the literature, recent examples of these applications can be seen in the multiple sclerosis (MS) disease using deep learning architectures. Salem et al. (2020) proposed a CNN-based architecture to detect new T2-w lesions to predict prognosis worsening. They proposed an architecture with two input channels (basal and follow-up images), where the first part of the Unet architecture learns the deformation fields (DFs) and image features, and in the second part of the architecture, another Unet performed the final detection and the segmentation of the new T2-w lesions. As another example, Gessert et al. (2020) proposed an attention-guided two-path CNN approach to detect the lesion activity in terms of new and enlarging lesions between two-time points, given basal and follow-up images as inputs.

In the hematoma expansion problem, similarly, we have basal and follow-up information. These two images can be used to detect the clinical worsening of the patient or the segmentation of the lesion growth in given two time points. To the best of our knowledge, there is no lesion growth detection algorithm established in the hematoma expansion literature.

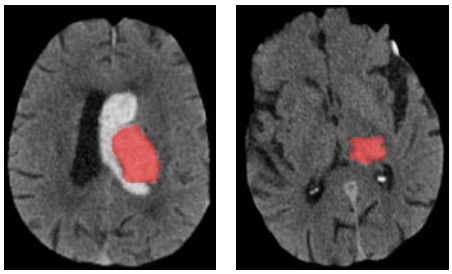
3. Material and methods

3.1. Dataset

The dataset used in the project was obtained at Dr. Josep Trueta's hospital in Girona, Spain. It consists of 206 cases, each with non-contrast head CT scans. The image examinations were performed on a 128-slice CT scanner (Philips Healthcare) and NCCT volumes had a slice thickness was 3 mm and a gap of 1.5 mm.

From now on, we will refer to the term "basal" for the initial scans, while "follow-up volumes" will denote subsequent scans performed after 24 hours.

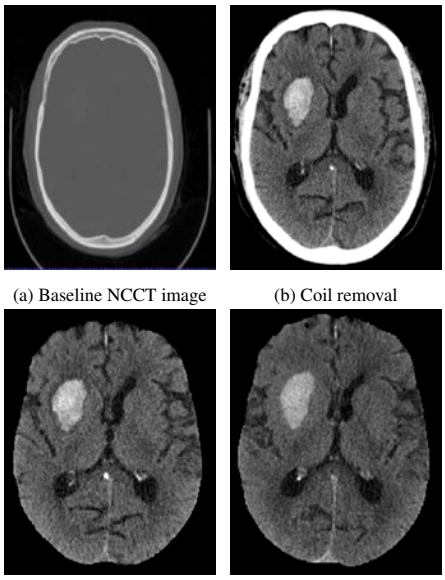
Our dataset includes 206 basal and follow-up cases, with 41 showing hematoma expansion and retaining 165 not showing hematoma expansion and with a mean growth of 3.03 ml and a standard deviation of 10.15 ml. In our study, hematoma expansion cases were referred to as positive cases. This subset of cases with hematoma expansion accounts for roughly 20% of the total dataset.



(a) IVH case with ground truth.

(b) IPH case with ground truth.

Figure 2: Example of IVH and IPH cases from the dataset.



(a) Baseline NCCT image

(b) Coil removal

(c) Baseline image preprocessed and registered.

(d) Follow-up image preprocessed

Figure 3: Preprocessing pipeline (a) Basal image NCCT (b) Basal image after coil removal (c) Basal image after pre-processing (d) Follow-up image after pre-processing.

3.1.1. Segmentation

According to the literature, several studies (Ma et al. (2022), Tingting et al. (2022), Teng et al. (2021)) have utilized segmentation masks within the prediction pipeline to establish the region of interest (ROI). In our study, we similarly incorporated segmentation masks for the images.

The segmentation masks for our dataset were obtained using 3D-based Unet architecture with squeeze-and-excitation blocks following the work of Abramova et al. (2021). Obtained segmentation masks were approved and refined when needed by an expert neurologist from Dr. Josep Trueta Hospital. In our study, hematoma volumes were calculated based on approved segmentation from the neurologist.

3.1.2. IVH and IPH

IVH, also known as intraventricular bleeding, is a type of hemorrhage that takes place within the brain ventricles where cerebrospinal fluid is produced, and it is an extension of hemorrhage within the brain parenchyma. The adjacent hemorrhagic stroke lesion near the ventricles is one possible origin of it. On the other hand, intraparenchymal hemorrhage (IPH) is the name of bleeding that only appears in the brain tissue. Since in the ground truth segmentation masks intraventricular hemorrhage was not delineated as a stroke class, it was not segmented. Figure 2 shows an example of accepted ground truths after neurologist examination.

In the Trueta dataset, we have 33 IVH cases (%16 of the dataset), where 5 cases are labeled as positive hematoma expansion. Since the hematoma expansion classification problem requires data to be well distinguished between positive and negative classes, we wanted to investigate the effect of the existence of IVH cases for this problem.

3.1.3. Data preparation

The initial pre-processing of NCCT images requires coil removal and skull stripping (see Figure 3) because regions outside of the brain may mislead the algorithm and result in undesirable outcomes. For the coil removal, the image was binarized, and the biggest connected component, in this case, the head, was kept. The skull stripping algorithm is performed by removing the borders using morphological operations and the final brain extraction is made based on the biggest connected components. The intensity ranges of the images were scaled from 0 to 90.

To enable analyses based on basal and follow-up images, a multi-level affine registration was performed. This involved an initial rigid registration, with the basal image selected as the moving image and the follow-up image as the fixed image. Figure 3 shows the preprocessing approach, which contains coil removal, intensity scaling, skull stripping, and registration.

3.1.4. Dataset balancing

In the field of deep learning, one common challenge that researchers and practitioners often face is imbalanced datasets, where the number of instances in one class significantly outweighs the other. This scenario can make training models difficult because the model is biased towards the majority class, resulting in poor performance of the minority class. To address this issue, various techniques such as undersampling, oversampling, and class weighting have been used, but they do not always produce satisfactory results.

To overcome the problem of class imbalance in our dataset, we decided to create a balanced dataset that not only equalized the class distribution but also took into account the different volumes of basal lesions within the positive class. In our dataset, the positive class accounted for only 20% of all instances, while the nega-

tive class accounted for the remaining 80%. This significant class imbalance made it difficult to train a model capable of capturing patterns and making accurate predictions for the positive class.

To address this issue, we focused on the basal volume of lesions, considering it a crucial factor in the classification task. We divided the positive class into different ranges based on the basal lesion volume. Within this specific volume range, we ensured that we had an equal number of cases from the negative class. This approach allowed us to create a balanced dataset not only in terms of class distribution but also in terms of lesion volume distribution. This approach aimed to enhance the model's ability to generalize and make accurate predictions for both the positive and negative classes, regardless of the basal lesion volume. During this stage, we took extra care to maintain the reliability of our analysis. We removed lesions from consideration if their initial basal volume was below 5 ml or above 60 ml. It's worth noting that these particular lesions were only found in the negative class. This step was crucial to prevent any biases that could have affected our results. Finally, we obtained a balanced dataset containing 70 patients, given in Table 1. The balanced dataset is used for the 2D and 3D detection and prediction models.

Table 1: Data distribution in the balanced dataset.

Range of basal volume, ml	Number of cases
5-10	28
10-20	16
20-30	8
30-40	6
40-60	12

3.2. Methodology

In this work, we worked on two different tasks. The first one is the detection of the hematoma growth, meaning that given basal and follow-up image information, performing classification that detects the cases with hematoma expansion. The second one is the prediction of the future hematoma growth occurrence through the analysis of solely basal image data. The general pipeline of these two tasks is given in Figure 4, where DNN corresponds to different DNN architectures implemented.

The detection model was designed to provide a quick and reliable deep learning-based approach for clinical settings. This model is aimed to define the critical deterioration of the patient in the following hours to the onset time. The condition for clinical worsening is the same as hematoma expansion, meaning that having more than 6 ml increase in the absolute volume or 33% relative volume increase. Clinical worsening can be detected by first having the segmentation masks and then calculating the difference between basal and follow-up

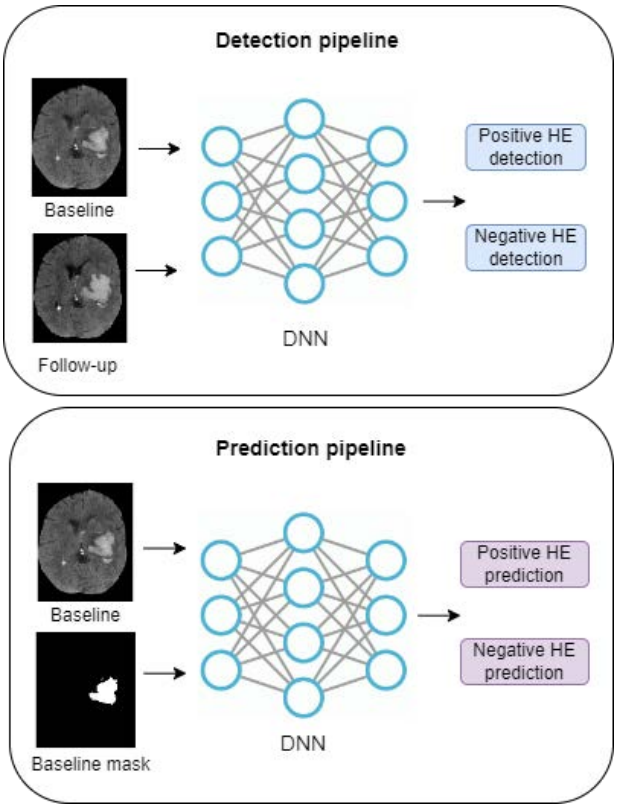


Figure 4: The proposed detection and prediction pipelines.

image masks. However, segmentation of the lesions is usually a semi-automatic and time-consuming approach considering the clinical settings. Recognizing the urgency and high mortality associated with this disease, we aimed to provide a deep learning approach capable of detecting differences between basal and follow-up image information without relying on segmentation masks. In this model, baseline and follow-up images were given as two-channel inputs to the DNN model.

The prediction model was designed to provide a deep learning based approach to predict the HE occurrence in the follow-up scan for a patient, using only basal image information. This model utilizes basal image features to forecast the clinical worsening of the patient at the next time point. The ground truth mask is used to provide attention to the lesion features itself, rather than other structures within the frame. In this model, baseline and baseline masks were given as two channel inputs to the DNN model.

In light of the existing literature that highlights the successful use of non-contrast computed tomography (NCCT) images for predicting hematoma expansion, both in two-dimensional (2D) and three-dimensional (3D) formats, our objective was to perform a comprehensive analysis using multidimensional imaging and compare these methods within a single dataset.

CT scans inherently capture data in the form of 3D volumes. However, for the purpose of training models, it is possible to adopt a 2D approach with these images. In this approach, the 3D volume is essentially considered a collection of 2D slices. This strategy provides computational efficiency by treating each slice as an independent image, making it appropriate for situations with limited data. Nevertheless, a drawback of the 2D approach is that it often disregards the 3D spatial information and inter-slice relationships present within the volume. As a consequence, important contextual details that could contribute to accurate analysis and interpretation may be overlooked or lost. On the other hand, 3D approaches keep the spatial information with the cost of expensive model training and less amount of data.

To initiate our investigation, we began by focusing on the 2D detection task, as it served as an initial baseline approach for addressing the prediction problem. As part of this task, we included each slice of the patient's NCCT scan in our dataset, specifically focusing on slices where the lesion was present and the lesion mask contained more than 100 pixels.

In a 2D slice-based approach, it's crucial to address the challenge of patient data distribution across sets. Ensuring the exclusive presence of a patient's data in a single set is vital to avoid data leakage, which can lead to unreliable and biased results. To avoid accidental information leakage, we carefully assigned slices to sets based on unique patient IDs. This method preserved data integrity during the training, validation, and testing stages. During the evaluation, we used two distinct metrics: image-wise and patient-wise. The image-wise evaluation assessed probabilities or predictions for individual slices independently, without considering relationships between slices from the same patient. In the patient-wise approach, we calculated the mean prediction value for a patient using all their slices.

The use of 2D detection analysis proved invaluable in understanding the most appropriate and effective models and implementations compatible with our dataset. In this stage, we experimented with three different groups of methods.

1. Modified Unet encoder architecture: Unet, Unet-SE and Unet-AG.
2. Transfer learning algorithms: Densenet121, EfficientNet-B0, Resnet18 and Resnet34.
3. Vision transformers: Swin-t and R50-Vit.

We experimented with 9 models belonging to given 3 groups in 2D detection and prediction problem. In 3D problem, because of the availability of given models in 3D, we implemented EfficientNet-B0, Densenet121, Resnet18, Unet, Unet-SE and Unet-AG.

Furthermore, we conducted a study to examine the impact of different scales of the same data on the output,

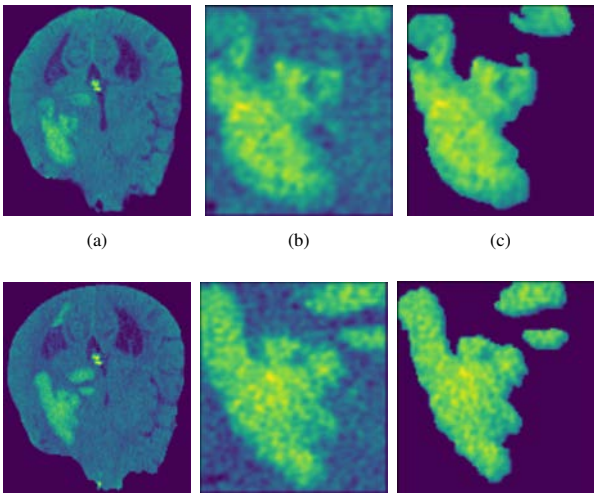


Figure 5: Different input types used in the study, (a)(b)(c) Basal image for the whole image, ROI and lesion based, (d)(e)(f) Follow-up image for the whole image, ROI and lesion based.

considering both ROI-based and whole image/volume-based approaches discussed in the literature. We investigated the following tests with varying scales of the data are given in Figure 5:

1. Whole image/volume-based: As the name suggests, this approach encompassed the entire raw slice or volume of the data. It considered the complete context provided by the original image or volume.
2. ROI-based: In this approach, a region of interest (ROI) was defined by considering the lesion information and the surrounding pixels within a bounding box. The analysis took into account the localized area surrounding the lesion.
3. Lesion-based: This approach involved isolating a specific lesion from the background information by applying a mask. The analysis focused solely on the lesion itself.

By examining these different scales, we aimed to gain insights into how the choice of scale impacts the output and performance of the methods under investigation. Following the 2D detection approach, we proceeded to evaluate the performance of the successful models within the 2D prediction model. Later, we repeated model tests for 3D detection and prediction problems.

3.3. Deep learning architectures

3.3.1. Unet encoder variations

Unet encoder. The Unet architecture was first proposed by Ronneberger et al. (2015) for the cell segmentation task. It was specifically designed for biomedical image segmentation tasks, but its effectiveness has been

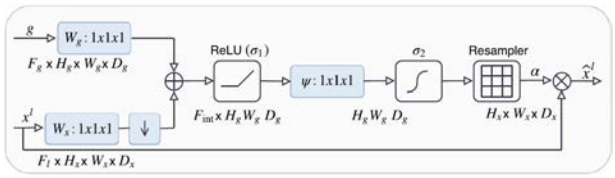


Figure 6: Schematic of the additive attention gate by Schlemper et al. (2019)

demonstrated in various other domains as well. It consists of a contracting path (encoder) and an expanding path (decoder), which form a U-shaped network. The encoder part of the Unet contains a series of convolutional layers with rectified linear unit (ReLU) activations, followed by max-pooling layers. These layers gradually reduce the input's spatial dimensions while increasing the number of feature channels, allowing the network to capture high-level abstract features. This ability to capture features makes the Unet encoder a powerful feature extractor that can be used in various tasks.

The Unet encoder architecture was implemented to a 3D HE prediction problem in the literature by (Zhong et al., 2021). This architecture includes a series of convolutional blocks followed by Instance normalization and Leaky Relu activation function, two max pooling operations and two global residual connections. In our study, we refer to this Unet-like encoder with residual connections as our base Unet encoder model. We implemented this model in 2D and 3D settings.

Unet-AG. An attention gate is a mechanism used in deep learning models, particularly in the context of image segmentation, to selectively focus on relevant regions or features while suppressing irrelevant or noisy information. It helps the model to attend to specific areas of an input image, enabling more accurate and refined segmentation results. Schlemper et al. (2019) proposed an additive attention gate approach (see Figure 6) that can be integrated into a standard CNN model, resulting in higher sensitivity and prediction accuracy. In their work, they proposed Attention-Unet for segmentation and AG-Sononet for the classification tasks.

The objective of employing the "attention" strategy is to learn to focus on target structures while disregarding irrelevant areas within an image. In the context of our detection problem, our intention was for the model to direct its attention toward focusing on the differences between the given 2 input images. Conversely, in our prediction problem, our aim was for the model to prioritize features that are relevant to the expansion occurrence. In our study, we introduced an attention-gated classification model from the base Unet encoder model with global residual connections given in Figure 7. The architecture takes given input images and implements a series of convolutional blocks. Each of these

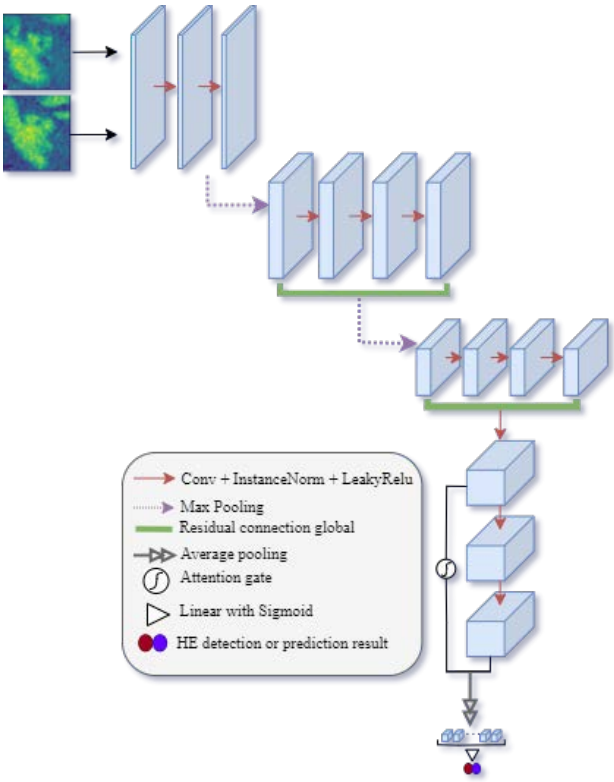


Figure 7: Schematic of proposed Unet encoder with Attention gate. Example of a detection model with input of 2 channels basal and follow-up ROI.

blocks contains a convolutional layer followed by instance batch normalization and Leaky Relu activation function. For the following two convolutional blocks, a global residual connection is added. An attention gate mechanism is added between the output of the 9th convolution and the last convolution output.

Unet-SE. The squeeze-and-excitation blocks were first introduced by (Hu et al., 2017). SE blocks can be used as building blocks for the current CNNs at an additional slight cost of computation. Their main goal is to enhance the representational power of a neural network. The SE block aims to explicitly model the interdependencies between channels in a feature map. It captures the importance of each channel and adaptively recalibrates them to improve the overall feature representation. The SE block consists of two main steps: squeeze and excitation.

1. Squeeze: In this step, the spatial dimensions of the feature map are reduced to capture global information about the channel interdependencies. It involves applying an average pooling operation over the spatial dimensions. This operation aggregates the feature maps from each channel, creating a channel descriptor.
2. Excitation: In this step, the channel descriptor obtained from the squeeze step is used to model the

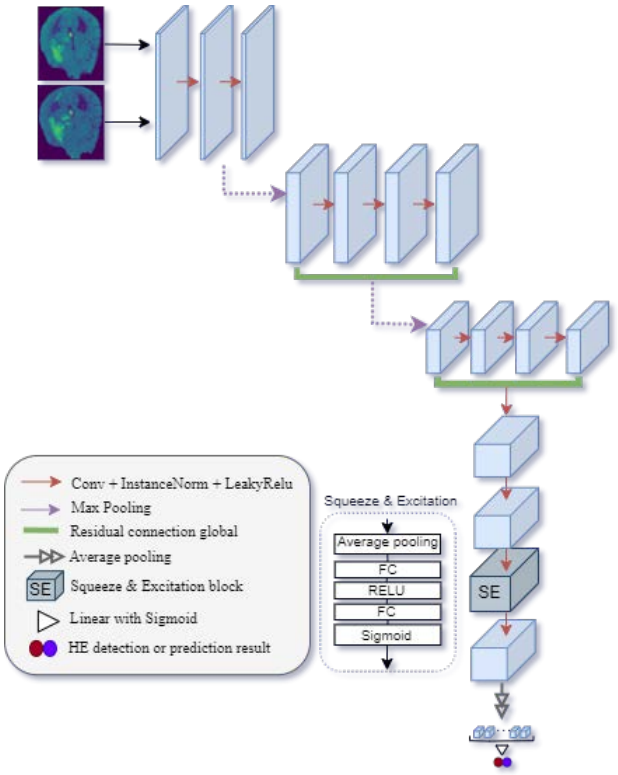


Figure 8: Schematic of proposed Unet-SE encoder with squeeze and excitation module. Example of a detection model with input of 2 channels basal and follow-up whole image.

interdependencies between channels and generate channel-wise importance scores. The channel descriptor is passed through a fully connected layer, reducing the dimensionality of the descriptor. It is followed by a ReLU activation function and another fully connected layer where the dimensionality of the descriptor is increased to the original dimension. Then a Sigmoid activation function is applied. Finally, the output of the excitation step is a set of channel-wise scaling factors that capture the importance of each channel.

The SE block is then applied to the original feature map by multiplying the original feature map with the channel-wise scaling factors. This recalibration operation dynamically adjusts the contribution of each channel, allowing the network to focus on more informative channels and suppress less useful ones. The schematic of the proposed Unet-SE architecture is given in Figure 8).

3.3.2. Vision transformers

The effectiveness of Transformers and attention mechanisms has been demonstrated in the field of natural language processing (NLP). Drawing inspiration from their achievements, several studies have attempted to apply transformer architectures in the domain of computer vision. Dosovitskiy et al. (2020) were the first ones to apply pure transformer architecture in computer vision. In our study, we implemented two vision transformer architectures: Swin (Liu et al., 2021), and R50-ViT (Chen et al., 2021).

In vision transformer architecture, an image is divided into non-overlapping patches each having a size of 16x16. Each patch represents a specific area of the image. To obtain a sequence of patch embeddings, these patches are flattened and linearly projected. Positional encodings are added to the patch embeddings to incorporate positional information into the model. The spatial location of each patch in the image is encoded by these encodings. The patch embeddings with positional encodings are then fed into a Transformer encoder which contains multiple layers, each containing a multi-head self-attention mechanism and a feed-forward neural network. The final patch embedding is used as the image's global representation. This representation is then fed into a classification head, which uses the image to make predictions, such as object recognition or image classification.

Although vision transformers typically demonstrate superior performance on large datasets and architectures, it is still feasible to discover smaller vision transformer architectures with a lower number of parameters, which can be pre-trained on ImageNet weights, on Hugging Face Transformers model hub (Wolf et al., 2020) and Pytorch torchvision package.

Swin transformer. Swin transformer is a Vision Transformer variant that addresses the original architecture's limitation in handling large images. It implements a hierarchical structure that divides the image into stages, allowing the model to effectively capture both local and global information. Rather than directly applying the self-attention mechanism on the patch embeddings, the Swin Transformer introduces a hierarchical structure of stages. Each stage processes a lower-resolution version of the image at a different scale. The shifted window mechanism is used within each stage of the Swin Transformer. Shifted windows are used to capture local dependencies instead of regular non-overlapping patches. This allows the model to handle large images without losing fine-grained details.

R50-ViT. TransUnet is a hybrid model including Resnet50 and ViT in the encoder and Unet in the decoder part Chen et al. (2021). This model is proposed for the multi-organ segmentation task and showed a successful performance. R50-ViT is the hybrid encoder part of this network combining the successful Resnet50 model with ViT.

3.3.3. Transfer learning models

Transfer learning has emerged as the state-of-the-art technique in the literature in the context of the hematoma expansion problem due to its efficiency and demonstrated success. In our study, we employed three architecture models: Densenet121 introduced by (Huang et al., 2018), EfficientNet B0 proposed by (Tan and Le, 2019), and Resnet architecture by (He et al., 2016).

3.4. Training and validation strategy

In our training, we utilized 5-fold cross-validation to ensure an unbiased and comprehensive evaluation of performance. During the network training, 75% and 15% of the data were assigned as training and test set, sequentially. A validation set is created from the training set. In the end, we derived a total of 54 patients for training, 6 for validation, and 10 for testing sets. For the 2D task, the training, validation, and test sets contained 1288, 160, and 227 slices, respectively. It's important to note that we maintained the same data distribution across these sets to ensure fairness and consistency in our evaluations.

3.5. Metrics

To evaluate the classification model's success we used Accuracy, F1 Score, Sensitivity, Specificity, and ROC AUC score. Accuracy measures the overall correctness of a classifier by calculating the ratio of correct predictions to the total number of predictions. It provides a general indication of how well the classifier performs across all classes, however, it is not suitable for imbalanced datasets. The accuracy formula is given by:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

Precision measures the proportion of correctly predicted positive instances (true positives) out of all instances predicted as positive (true positives and false positives). The precision formula is given by:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

Recall, also known as sensitivity or true positive rate, measures the proportion of correctly predicted positive instances (true positives) out of all actual positive instances (true positives and false negatives). The recall formula is given by:

$$\text{Recall (Sensitivity)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

Precision and recall metrics are not directly used in our evaluation metrics but they were used to calculate the F1 Score, which is a harmonic mean of precision and recall, offering a balanced assessment of both metrics. The harmonic mean used in the F1 score calculation places more weight on low values, meaning that the F1 score will be lower if either precision or recall is low. This makes the F1 score a suitable metric when there is an imbalance between the positive and negative classes in the dataset. The F1 score formula is given by:

$$\text{F1 score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

Specificity, also known as true negative rate, measures the proportion of correctly predicted negative instances (true negatives) out of all actual negative instances (true negatives and false positives).

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{5}$$

The Receiver Operating Characteristic (ROC) curve is a graphical representation of a binary classification model's performance. At various classification thresholds, it shows the trade-off between the true positive rate (TPR) and the false positive rate (FPR). The Area Under the ROC Curve (AUC) is a metric that quantifies the model's overall performance across all classification thresholds. During the parameter tuning and model evaluation decisions for the validation set, the ROC AUC scores were used in the decision process.

3.6. Data augmentation

In the context of 2D detection and prediction, we incorporated random horizontal flipping and applied affine transformation. For 3D detection and prediction, we utilized a random affine transformation in three-dimensional space. The affine transformation parameters were as follows: a rotation angle of 10 degrees, horizontal and vertical translations of 10 each, and a scaling factor of 1.2. Furthermore, the inputs were normalized between 0 to 1 using min-max normalization.

3.7. Implementation details

The project was implemented using the open-source deep learning framework PyTorch. The deep learning algorithms were implemented from PyTorch's torchvision library, Hugging Face's Transformers model hub (Wolf et al., 2020), and models from MONAI (Cardoso et al., 2022). The experiments were implemented using Pytorch 2.0.1 and CUDA 11.7 on a Linux environment. The 2D training was performed on an Nvidia Titan V GPU with 12 GB of memory. The 3D experiments were implemented on an Nvidia A30 Tensor core GPU with 24 GB of memory.

4. Results

This section contains the results of the different experiments mentioned in the Subsection 3.2 Methodology. The results section will be split into two subsections to cover detection and prediction experiments. Furthermore, each subsection will be divided into two to cover 2D and 3D experiments. The results were calculated based on average test scores of 5-fold cross-validation. For the comparison, they were ordered based on their ROC-AUC scores.

4.1. Detection experiments

In the detection section, our aim is to introduce a quick deep learning-based approach to detect the patient worsening. Therefore, in this section, we experimented with whole image/volume and ROI-based analyses.

4.1.1. 2D detection results

For the 2D experiments, we used a batch size of 4, a learning rate of 10^{-5} with Adam optimizer. We train the models for 25 epochs with focal loss where StepLR scheduler is applied with the step size of 15 and gamma of 0.1.

In our 2D experiments, we utilized a thresholding approach to select the appropriate slices. This method involved selecting slices based on the size of the lesion within a given basal mask for each specific slice. If the basal mask contained more than 100 pixels, indicating that the lesion was sufficiently large for inclusion in our training, the slice was chosen. This procedure was applied to all dataset.

The selection of the threshold number involves a trade-off. A higher threshold resulted in larger and clearer lesions for the model. However, this also led to the elimination of more slices from the dataset, reducing the overall data size. We conducted experiments using various thresholds, including 50, 100, 200, 400, and 700. After evaluating performance, we obtained the optimized performance with the threshold of 100.

During the training phase, we conducted experiments using two different approaches: utilizing pre-trained weights from ImageNet and training from scratch. After evaluating the results, we found that training from scratch yielded superior outcomes compared to using pre-trained weights from ImageNet. We trained Densenet121 (7M), EfficientNetB0 (5M), Resnet34 (21M), Unet (7M), Unet-AG (8M) and Unet-SE (7M) models from scratch.

For the transformer training, we employed the Swin tiny (28M) and R50-ViT (99M) pre-trained architectures. In R50-ViT, we froze the ResNet50 backbone and fine-tuned the ViT component, which consisted of 7M parameters. In the Swin tiny architecture, we froze the backbone and fine-tuned the last PatchMerging and the following two SwinTransformerBlock. This architecture consisted of 15M parameters.

In the ROI-based experiments, we utilized the ROI of the basal and follow-up images belonging to the same slice number as the model inputs. We extracted the ROIs using the ground truth masks and subsequently normalized each of them using min-max normalization. This normalization process ensured that the intensity values of each ROI image were normalized within the range of 0 to 1. We performed zero padding to ensure that all ROIs has the same shape, 512x512. However, for the implementation of the R50-ViT model, images are resized to be 384x384, as stated in its documentation.

It is important to point out that, the ROIs that are used to train the network are obtained using ground truth masks. However, our main idea is to introduce a trained tool that can be used in clinical settings given the cropped ROIs. These ROIs can be obtained by visualization software such as 3D Slicer.

In the whole image experiments, similar to ROI experiments, basal, and follow-up slices were given as two-channel inputs into models, image settings are visualized in Figure 5. We performed min-max normalization and the shape of the slices was 512x512.

The 2D detection results calculated using whole images were given in Table 2a. In this case, Swin_t architecture showed the best performance achieving a 0.91 ROC-AUC score. The 2D detection results calculated using ROI images were given in Table 2b. It showed that the Unet model performed the best achieving a 0.92 ROC-AUC score.

This experiment concludes that it is possible to achieve almost the same high performance using ROI and whole images with different architectures in the case of a 2D detection problem. However, when also considering the architectures that rank second and third, ROI-based image results have a better general performance.

4.1.2. 3D detection results

For 3D experiments, we used a batch size of 2, a learning rate of 10^{-5} with Adam optimizer. We train the models for 25 epochs with binary cross entropy loss where the StepLR scheduler is applied with the step size of 15 and gamma of 0.1. During the training phase, we trained 7 models in 3D. These models were Densenet121, EfficienNetB0, Resnet18, Resnet34, Unet, Unet-AG and Unet-SE.

In the ROI-based and lesion-based experiments, after obtaining the cropped volumes, we performed min-max normalization and zero padding to size of 64x320x320, based on the greatest volume size of the lesion and ROI in the dataset. During the data preparation of the 3D

Table 2: Detection results.

(a) Top three test set average 2D detection results from 5-fold cross-validation, using the whole image, calculated patient-wise.

Model	Accuracy	F1-Score	Sensitivity	Specificity	ROC-AUC
Swin_t	0.91 (0.07)	0.90 (0.09)	0.87 (0.15)	0.96 (0.07)	0.91 (0.07)
EfficientNet	0.86 (0.10)	0.82 (0.14)	0.75 (0.23)	0.96 (0.07)	0.86 (0.10)
R50-ViT	0.85 (0.10)	0.82 (0.14)	0.76 (0.23)	0.96 (0.07)	0.86 (0.10)

(b) Top three test set average 2D detection from 5-fold cross-validation results using ROIs, calculated patient-wise.

Model	Accuracy	F1-Score	Sensitivity	Specificity	ROC-AUC
Unet	0.91 (0.07)	0.90 (0.09)	0.87 (0.15)	0.96 (0.07)	0.92 (0.03)
Unet-SE	0.91 (0.07)	0.90 (0.09)	0.83 (0.14)	1.00 (0.00)	0.91 (0.07)
Unet-AG	0.90 (0.08)	0.87 (0.11)	0.80 (0.17)	1.00 (0.00)	0.90 (0.08)

(c) Top three test set average 3D detection results from 5-fold cross-validation using the whole volume.

Model	Accuracy	F1-Score	Sensitivity	Specificity	ROC-AUC
Unet-SE	0.73 (0.20)	0.70 (0.24)	0.68 (0.27)	0.80 (0.30)	0.74 (0.20)
Unet-AG	0.71 (0.15)	0.65 (0.23)	0.60 (0.29)	0.83 (0.24)	0.72 (0.16)
Unet	0.71 (0.19)	0.68 (0.22)	0.68 (0.27)	0.75 (0.29)	0.72 (0.19)

(d) Top three test set average 3D detection results from 5-fold cross-validation using ROIs volume.

Model	Accuracy	F1-Score	Sensitivity	Specificity	ROC-AUC
Unet	0.80 (0.18)	0.75 (0.26)	0.71 (0.29)	0.87 (0.09)	0.80 (0.18)
Unet-SE	0.73 (0.20)	0.70 (0.24)	0.68 (0.27)	0.80 (0.30)	0.74 (0.20)
Unet-AG	0.73 (0.18)	0.72 (0.18)	0.71 (0.27)	0.75 (0.38)	0.74 (0.18)

dataset, we needed to perform specific arrangements because of the GPU memory limitation. We empirically found that the optimum setting was to resize the volumes and have the batch size of 2. Therefore, the input volumes were resized to 32x256x256.

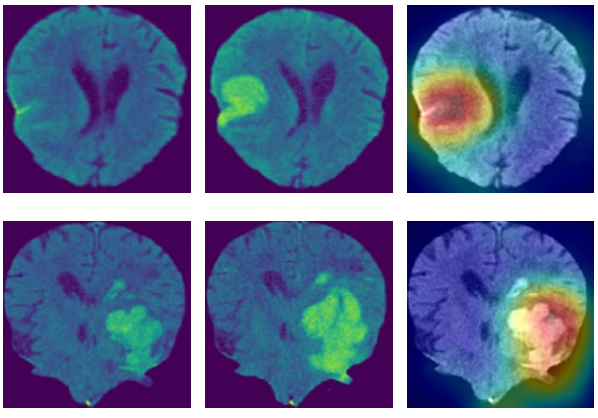
In the 3D detection results using whole volumes, the Unet-SE model performed the best achieving a 0.74 ROC-AUC score given in Table 2c. In terms of ROI results, the Unet model showed superior performance by having a ROC-AUC score of 0.80 given in Table 2d.

The experiments conclude that for the 3D detection model, ROI volume performed better than the whole image approach. Furthermore, overall in detection models 2D detection model performed better than the 3D detection model.

4.2. Model explainability

Grad-CAM (Gradient-weighted Class Activation Mapping) is a technique to visualize and understand the reasoning behind the predictions made by convolutional neural networks (CNNs), introduced by (Selvaraju et al., 2017). It generates a heat map that highlights the regions of the input image that were crucial in determining the prediction. This heat map is created by leveraging the gradients of the target class with respect to the convolutional feature maps.

The Grad-CAM results based on our 2D detection model showed that the detection model is making decisions based on the difference between provided two inputs. In the end, it determines the HE by focusing on the lesion area itself as given in Figure 9.



(a) (b) (c)

Figure 9: (a) Basal image, (b) Follow-up image, (c) Grad-CAM obtained on the basal image, using 2d detection model, where the red areas represent the regions of the input image that are highly important for the model's prediction.

4.3. Prediction experiments

In the prediction section, our aim is to predict patient prognosis only based on basal image and mask information. The model received the basal image and basal mask as inputs, which were provided as two-channel inputs. This two-channel application aims to have attention on the lesion. In this section, we experimented with whole image/volume, ROI-based and lesion-based analyses, we obtained our results from 5-fold cross-validation.

4.3.1. 2D prediction results

The 2D prediction model settings are the same as 2D detection models, including settings for obtaining ROI-based and whole images. They are explained in the subsection 4.1.1 2D detection results. In the lesion-based experiments, we obtained the input images by masking the basal image with the basal mask. Similar to ROI-based experiments, we performed a zero padding to ensure that all lesion-based basal images has the same shape, 512x512. We performed min-max normalization for each lesion-based image.

The 2D prediction results that are calculated using whole images were given in the Table 3a. Results show that EfficientNetB0 model performed the best achieving the ROC-AUC score of 0.720. Considering the complexity of the prediction problem compared to the detection problem, we obtained overall lower results, as expected.

For the ROI results, Resnet34 model performed the best, achieving 0.695 ROC-AUC score. The ROI based 2D prediction results are given in the Table 3b. The R50-ViT model demonstrated the highest performance among the lesion-based image results, with a ROC-AUC score of 0.690, given in the Table 3c.

In conclusion, the 2D prediction results showed that the best setting for the prediction was with the whole-volume image setting with the EfficientNetB0 model.

4.3.2. 3D prediction results

The 3D prediction model settings are the same as 3D detection models, including settings for obtaining whole volume, ROI-based volume and lesion-based volume procedures. Considering the complexity of the prediction problem and limitations from 3D setting, with our current setting we could not obtain a stable prediction model. In the best setting, we obtained an average test ROC-AUC score of 0.559 using Resnet34 model.

4.4. Effect of patient-wise and image-wise evaluation metrics

During our experiments in 2D, we implemented patient-wise and image-wise evaluation metrics. We also performed different voting approaches on the assignment of the patient-wise metric. These voting methods contained three approaches. First, taking the mean of the probabilities belonging the same patient's slices and then assigning it to the patient probability. Second, taking the medium slice's probability value to be the probability of the patient. Lastly, taking the maximum probability among the slices of a patient and assigning it to be the patient probability.

Our results showed that mean voting approach gave the best results. Furthermore, patient-wise approach outperformed to image-wise in almost all model experiments. This concludes that, patient-wise patient probability assignment is more reliable.

4.5. Effect of IVH cases

As mentioned in the subsection 3.1.2 IVH and IPH, we investigated the effect of excluding the IVH cases. The number of IVH cases in the dataset was 10 in total of 70 cases. Therefore removing these cases came with a trade off of having a much smaller dataset but containing only IPH cases. For some models we observed a positive effect when IVH cases were excluded. However for some others, we had the opposite. This phenomena can be explained by the way of different architectures process the data. However, at this stage, considering the bias coming from having a much smaller dataset, the specific factors contributing to this occurrence remain unexplained.

5. Discussion

In this work, we investigated hematoma expansion detection and prediction, analysing the use of 2D and 3D deep learning strategies and also the use of different image input approaches. The applications and interpretations of the work can be divided into two subsections.

5.1. Detection approaches

The 2D detection approaches presented in this work are the first attempts to implement a lesion growth detection model specifically for the problem of hematoma expansion. This success can be attributed to the relatively lower difficulty of the detection task compared to the prediction task, as well as the slice-by-slice nature of 2D models, which proved to be beneficial in our small dataset. Patient-wise evaluation metrics were defined and utilized in our experiments, with the average voting metric demonstrating success. Remarkable results were achieved using both whole image and ROI-based input types, indicating that these two image types can be effectively utilized in the 2D detection problem. Overall, we obtained a 0.91 ROC-AUC value using the Swin transformer with whole images and a 0.92 ROC-AUC score using Unet with the ROI images, shown in the Tables 2a and 2b. The presented approach was supported by the Grad-CAM explainability model, visualizing that the model was making decisions based on the difference between lesions in given two input images.

Regarding the 3D approaches, it comes with the advantage of having spatial information, but it also has the disadvantage of having a smaller number of samples to train. In this model, we obtained the best value with a 0.80 ROC-AUC value with the Unet model and ROI-based volume setting, given in the Table 2d.

It concludes that for the detection problem, ROI-based image setting showed a successful performance,

Table 3: Prediction results.

(a) Top three test set average 2D prediction results s from 5-fold cross-validation using the whole image, calculated patient-wise.

Model	Accuracy	F1-Score	Sensitivity	Specificity	ROC-AUC
EfficientNet	0.71 (0.19)	0.70 (0.19)	0.68 (0.20)	0.76 (0.23)	0.72 (0.19)
Densenet	0.63 (0.16)	0.62 (0.20)	0.68 (0.27)	0.60 (0.21)	0.64 (0.13)
Unet-SE	0.63 (0.14)	0.52 (0.20)	0.51 (0.32)	0.75 (0.15)	0.64 (0.10)

(b) Top three test set average 2D prediction results s from 5-fold cross-validation using ROIs, calculated patient-wise.

Model	Accuracy	F1-Score	Sensitivity	Specificity	ROC-AUC
Resnet34	0.68 (0.16)	0.65 (0.22)	0.64 (0.31)	0.75 (0.22)	0.69 (0.15)
R50-ViT	0.68 (0.12)	0.74 (0.09)	0.80 (0.17)	0.55 (0.33)	0.67 (0.14)
Unet-SE	0.66 (0.21)	0.63 (0.25)	0.60 (0.28)	0.75 (0.15)	0.67 (0.20)

(c) The test set average 2D prediction results from 5-fold cross-validation using the lesion-based image, calculated patient-wise.

Model	Accuracy	F1-Score	Sensitivity	Specificity	ROC-AUC
R50-ViT	0.68 (0.12)	0.67 (0.19)	0.67 (0.27)	0.69 (0.18)	0.69 (0.11)
Densenet	0.71 (0.15)	0.78 (0.09)	0.87 (0.09)	0.50 (0.35)	0.69 (0.16)
Unet-AG	0.60 (0.15)	0.52 (0.32)	0.60 (0.41)	0.60 (0.33)	0.60 (0.13)

considering that the model is actually focusing on the difference between given two lesion inputs, putting hard attention on the lesion area removes other distractions and helps with the detection process. The 3D model achieved lower ROC-AUC scores compared to the 2D detection model. This difference could be attributed to the smaller number of samples used during the training of the 3D model. To enhance the performance of the 3D model, it may be beneficial to increase the training sample size or explore other techniques to address the limitations.

5.2. Prediction approaches

For the 2D prediction approaches, In the hematoma expansion prediction problem, we observed that the whole image setting outperformed ROI and lesion-based images. This is the opposite of the 3D detection conclusion. However, the nature of the prediction model is different than the detection model, therefore the information needed to make a decision is different. The success of the whole images over ROI-based images might be related to the positional information of the given lesion, which is lost if the input is ROI or lesion-based images. In this model, the best setting was achieved by the EffiecientNetB0 model with a ROC-AUC score of 0.72, given in the Table 3a. Despite the prediction problem being considered more challenging, the slice-by-slice approach enabled us to train with more data, leading to a promising result.

Regarding the 3D prediction, being the most challenging problem in this study, we believe that challenges occurred because of the complexity of the prediction model and less number of training samples for the 3D training.

Overall, the results obtained from the 2D detection approaches were the most successful ones among all the experiments conducted. For the detection approaches, ROI-based image setting and for the 2D prediction problem, whole image setting showed the best performance. With the differences (standard deviation) present in the dataset, when we performed the dependent t-test for paired samples for the statistical analysis between approaches we observed that the results are not significantly conclusive.

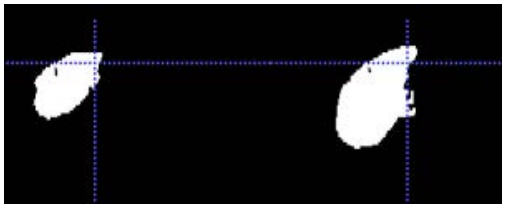
5.3. Limitations and future work

Throughout this study, our primary constraint revolved around the limited size and imbalanced nature of the available data. Given the complex feature characteristics and intraclass heterogeneity, especially constructing an effective prediction model requires the utilization of robust representations and much larger dataset to obtain a more reliable generalization of the problem.

Another limitation was the lack of spatial information in the 2D models. Even though in the 2D detection and prediction models we achieved promising results, we observed some misclassified cases because of the lack of spatial information in these models. In the 2D detection model, we observed a misclassification that happened when the lesion growth is in the sagittal plane instead of the axial plane. Since the slice view is taken from the axial view, basal and follow-up images look alike for most of the slices of the lesion, leading an average probability score of having a non-HE detection result, even though the right label is HE. The case is shown in Figure 10.

The 3D model training is the most appropriate way of using the spatial information of the 3D volumes, but it comes with the limitations of having fewer samples, especially if the dataset is small. In future work, we would like to replicate the experiments with a greater number of samples within a multicenter study.

(a) Axial view



(b) Sagittal view

Figure 10: The lack of spatial information observed in 2D detection problem.

6. Conclusions

In this work, we implemented and analysed a set of approaches to tackle two different problems related to the clinical management of the patients with hemorrhagic stroke. The first approach focused on detecting hematoma lesion growth in longitudinal NCCT images, from basal and follow up scans of a patient. The objective was to develop an automated deep learning pipeline that could accurately measure the progression of the hematoma over time and detect clinical worsening. This approach served as a foundation for the second approach.

The second approach was a challenging prediction model, particularly designed to classify cases of the future HE occurrences. Unlike the detection model, this prediction model relied only on basal image information to make its classification. The main purpose of this model was to identify features within the baseline images that were highly correlated with the likelihood of future HE.

We studied the impact of using 2D vs 3D strategies and also the use of different inputs to the model (the whole brain image, a ROI-based including the hematoma and the lesion-masked image of the case). We obtained promising results for the 2D, 3D detection, and 2D prediction tasks. For the 2D detection task, we obtained the ROC-AUC value of 0.92 with the Unet model with an ROI-based image. For the 3D detection task, we obtained a 0.80 ROC-AUC value with the Unet model with ROI-based volume. The results suggest that using HE detection approach holds promise for further exploration in clinical settings.

Despite the challenge of the prediction task, we obtained a 0.72 ROC-AUC value with the EfficientNet-B0 model with the whole image setting in 2D. However, due to our data size limitation, we could not obtain a stable 3D prediction model. The results of this study demonstrated the potential of integrating deep learning into clinical practice, particularly for the early detection of HE. This technology could serve as a valuable tool to support clinicians in making timely decisions for patient care.

Acknowledgments

I would like to thank my supervisors Dr. Xavier Lladó and Dr. Arnau Oliver for their valuable supervision, support, and feedback. Special thanks to Valeriia Abramova for her help, especially during the data collection and preparation steps and her recommendations. I would like to thank Hospital Dr. Josep Trueta, especially Dra. Yolanda Silva, for providing the dataset for research.

I am sincerely thankful to the MAIA program, professors, and organizers for their influence and inspiration on this remarkable journey of knowledge. Lastly, heartfelt thanks to my family and loved ones for their unconditional love, encouragement, and support throughout this academic journey.

References

Abramova, V., Clèrigues, A., Quiles, A., Figueredo, D.G., Silva, Y., Pedraza, S., Oliver, A., Lladó, X., 2021. Hemorrhagic stroke lesion segmentation using a 3d u-net with squeeze-and-excitation blocks. Computerized Medical Imaging and Graphics 90, 101908. doi:10.1016/j.compmedimag.2021.101908.

Cai, J., Zhu, H., Yang, D., Yang, R., Zhao, X., Zhou, J., Gao, P., 2020. Accuracy of imaging markers on noncontrast computed tomography in predicting intracerebral hemorrhage expansion. Neurological Research 42, 973–979. doi:10.1080/01616412.2020.1795577.

Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murray, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., He, Y., Xu, Z., Hatamizadeh, A., Zhu, W., Liu, Y., Zheng, M., Tang, Y., Yang, I., Zephyr, M., Hashemian, B., Alle, S., Zalbagi Darestani, M., Budd, C., Modat, M., Vercauteren, T., Wang, G., Li, Y., Hu, Y., Fu, Y., Gorman, B., Johnson, H., Genereaux, B., Erdal, B.S., Gupta, V., Diaz-Pinto, A., Dourson, A., Maier-Hein, L., Jaeger, P.F., Baumgartner, M., Kalpathy-Cramer, J., Flores, M., Kirby, J., Cooper, L.A., Roth, H.R., Xu, D., Bericat, D., Floca, R., Zhou, S.K., Shuaib, H., Farahani, K., Maier-Hein, K.H., Aylward, S., Dogra, P., Ourselin, S., Feng, A., 2022. MONAI: An open-source framework for deep learning in healthcare. N/A doi:https://doi.org/10.48550/arXiv.2211.02701.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021. Transunet: Transformers make strong encoders for medical image segmentation. CoRR abs/2102.04306.

Chen, Y., Qin, C., Chang, J., Lyu, Y., Zhang, Q., Ye, Z., Li, Z., Tian, F., Ma, W., Wei, J., et al., 2022. A machine learning approach for predicting perihematomal edema expansion in patients with intracerebral hemorrhage. European Radiology doi:10.1007/s00330-022-09311-3.

Davis, S.M., Broderick, J., Hennerici, M., Brun, N.C., Diringer, M.N., Mayer, S.A., Begtrup, K., Steiner, T., 2006. Hematoma growth is a determinant of mortality and poor outcome after intracerebral hemorrhage. Neurology 66, 1175–1181. doi:10.1212/01.wnl.0000208408.98482.99.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An image is worth

16x16 words: Transformers for image recognition at scale. CoRR abs/2010.11929.

Dowlatshahi, D., Demchuk, A.M., Flaherty, M.L., Ali, M., Lyden, P.L., Smith, E.E., 2011. Defining hematoma expansion in intracerebral hemorrhage: Relationship with patient outcomes. Neurology 76, 1238–1244. doi:10.1212/wnl.0b013e3182143317.

Duan, C., Liu, F., Gao, S., Zhao, J., Niu, L., Li, N., Liu, S., Wang, G., Zhou, X., Ren, Y., et al., 2021. Comparison of radiomic models based on different machine learning methods for predicting intracerebral hemorrhage expansion. Clinical Neuroradiology 32, 215–223. doi:10.1007/s00062-021-01040-2.

Gessert, N., Krüger, J., Opfer, R., Ostwaldt, A.C., Manogaran, P., Kitzler, H.H., Schippling, S., Schlaefer, A., 2020. Multiple sclerosis lesion activity segmentation with attention-guided two-path cnns. Computerized Medical Imaging and Graphics 84, 101772. doi:10.1016/j.compmedimag.2020.101772.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) doi:10.1109/cvpr.2016.90.

Hillal, A., Ullberg, T., Ramgren, B., Wassélius, J., 2022. Computed tomography in acute intracerebral hemorrhage: Neuroimaging predictors of hematoma expansion and outcome. Insights into Imaging 13. doi:10.1186/s13244-022-01309-1.

Hu, J., Shen, L., Sun, G., 2017. Squeeze-and-excitation networks. CoRR abs/1709.01507.

Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q., 2018. Densely connected convolutional networks. CVPR 2017 .

Li, Y.L., Chen, C., Zhang, L.J., Zheng, Y.N., Lv, X.N., Zhao, L.B., Li, Q., Lv, F.J., 2023. Prediction of early perihematomal edema expansion based on noncontrast computed tomography radiomics and machine learning in intracerebral hemorrhage. World Neurosurgery doi:10.1016/j.wneu.2023.03.066.

Liu, J., Xu, H., Chen, Q., Zhang, T., Sheng, W., Huang, Q., Song, J., Huang, D., Lan, L., Li, Y., et al., 2019. Prediction of hematoma expansion in spontaneous intracerebral hemorrhage using support vector machine. EBioMedicine 43, 454–459. doi:10.1016/j.ebiom.2019.04.040.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. CoRR abs/2103.14030.

Ma, C., Wang, L., Gao, C., Liu, D., Yang, K., Meng, Z., Liang, S., Zhang, Y., Wang, G., 2022. Automatic and efficient prediction of hematoma expansion in patients with hypertensive intracerebral hemorrhage using deep learning based on ct images. Journal of Personalized Medicine 12, 779. doi:10.3390/jpm12050779.

Qureshi, A.I., Mendelow, A.D., Hanley, D.F., 2009. Intracerebral haemorrhage. The Lancet 373, 1632–1644. doi:10.1016/s0140-6736(09)60371-8.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. CoRR abs/1505.04597.

Rost, N.S., Smith, E.E., Chang, Y., Snider, R.W., Chanderraj, R., Schwab, K., FitzMaurice, E., Wendell, L., Goldstein, J.N., Greenberg, S.M., et al., 2008. Prediction of functional outcome in patients with primary intracerebral hemorrhage. Stroke 39, 2304–2309. doi:10.1161/strokeaha.107.512202.

Salem, M., Valverde, S., Cabezas, M., Pareto, D., Oliver, A., Salvi, J., Rovira, A., Lladó, X., 2020. A fully convolutional neural network for new t2-w lesion detection in multiple sclerosis. NeuroImage: Clinical 25, 102149. doi:10.1016/j.nicl.2019.102149.

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D., 2019. Attention gated networks: Learning to leverage salient regions in medical images. Medical Image Analysis 53, 197–207. doi:10.1016/j.media.2019.01.012.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. 2017 IEEE International Conference on Computer Vision (ICCV) doi:10.1109/iccv.2017.74.

Steiner, T., Diringer, M.N., Schneider, D., Mayer, S.A., Begtrup, K., Broderick, J., Skolnick, B.E., Davis, S.M., 2006. Dynamics of intraventricular hemorrhage in patients with spontaneous intracerebral hemorrhage: Risk factors, clinical impact, and effect of hemostatic therapy with recombinant activated factor vii. Neurosurgery 59, 767–774. doi:10.1227/01.neu.0000232837.34992.32.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. International Conference on Machine Learning, 2019 .

Tang, Z.R., Chen, Y., Hu, R., Wang, H., 2022. Predicting hematoma expansion in intracerebral hemorrhage from brain ct scans via k-nearest neighbors matting and deep residual network. Biomedical Signal Processing and Control 76, 103656. doi:10.1016/j.bspc.2022.103656.

Teng, L., Ren, Q., Zhang, P., Wu, Z., Guo, W., Ren, T., 2021. Artificial intelligence can effectively predict early hematoma expansion of intracerebral hemorrhage analyzing noncontrast computed tomography image. Frontiers in Aging Neuroscience 13. doi:10.3389/fnagi.2021.632138.

Tingting, Z., Kailun, C., Yiqing, L., 2022. Machine learning-based prediction study of hematoma enlargement in patients with cerebral hemorrhage. Journal of Sensors 2022, 1–7. doi:10.1155/2022/4470134.

Tsao., C.W., 2023. Correction to: Heart disease and stroke statistics—2023 update: A report from the american heart association. Circulation 147. doi:10.1161/cir.0000000000001137.

Wada, R., Aviv, R.I., Fox, A.J., Sahlas, D.J., Gladstone, D.J., Tomlinson, G., Symons, S.P., 2007. Ct angiography "spot sign" predicts hematoma expansion in acute intracerebral hemorrhage. Stroke 38, 1257–1262. doi:10.1161/01.str.0000259633.59404.f3.

Wan, Q., Kuang, Z., Deng, X., Yu, L., 2022. Bgsnet: Bidirectional-guided semi-3d network for prediction of hematoma expansion. 2022 IEEE International Conference on Image Processing (ICIP) doi:10.1109/icip46576.2022.9897794.

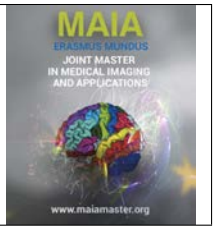
Wang, C., Deng, X., Yu, L., Kuang, Z., Ma, H., Hua, Y., Liang, B., 2021. Data fusion framework for the prediction of early hematoma expansion based on cnn. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) doi:10.1109/isbi48211.2021.9434043.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M., 2020. Transformers: State-of-the-Art Natural Language Processing. N/A , 38–45URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Zhong, J.w., Jin, Y.j., Song, Z.j., Lin, B., Lu, X.h., Chen, F., Tong, L.s., 2021. Deep learning for automatically predicting early haematoma expansion in chinese patients. Stroke and Vascular Neurology 6, 610–614. doi:10.1136/svn-2020-000647.

Medical Imaging and Applications

Master Thesis, June 2023



SYNCS: Synthetic Data and Contrastive Self-Supervised Training for Central Sulcus Segmentation

Vladyslav Zalevskyi, Kristoffer Hougaard Madsen

Danish Research Centre for Magnetic Resonance (DRCMR), Hvidovre, Denmark

Abstract

Bipolar disorder (BD) and schizophrenia (SZ) are severe mental disorders that have a significant impact on individuals and society. Early identification of risk markers for these diseases is crucial for understanding their progression and enabling preventive interventions. The Danish High Risk and Resilience Study (VIA) is a longitudinal cohort study that aims to gain insights into the early disease processes of SZ and BD, particularly in children with familial high risk (FHR). Understanding structural brain changes associated with these diseases during early stages is essential for effective interventions. The central sulcus (CS) is a prominent brain landmark related to brain regions involved in motor and sensory processing. Analyzing CS morphology can provide valuable insights into neurodevelopmental abnormalities in the FHR group. However, CS segmentation presents challenges due to its high morphological variability and complex shape, which are especially apparent in the adolescent cohort. This study explores two novel approaches for training robust and adaptable CS segmentation models that address these challenges. Firstly, we utilize synthetic data generation to model the morphological variability of the CS, adapting SynthSeg's generative model to our problem. Secondly, we employ self-supervised pre-training and multi-task learning to adjust the segmentation models to new subject cohorts by learning relevant feature representations of the cortex shape. These approaches aim to overcome limited data availability and enable reliable CS segmentation performance on diverse populations, removing the need for extensive and error-prone post- and pre-processing steps. By leveraging synthetic data and self-supervised learning, this research demonstrates how recent advancements in training robust and generalizable deep learning models can help overcome problems hindering the deployment of DL medical imaging solutions. Although our evaluation showed only a moderate improvement in performance metrics, we emphasize the significant potential of the methods explored to advance CS segmentation and their importance in facilitating early detection and intervention strategies for SZ and BD.

Keywords: segmentation, central sulcus, synthetic data, SynthSeg, self-supervised training, SimCLR, U-Net, multi-task learning

1. Introduction

1.1. Background

Bipolar disorder (BD) and schizophrenia (SZ) are severe mental disorders that impact approximately 0.7% and 1.0% of the population respectively (Robinson and Bergen, 2021). These conditions impose a significant burden on both individuals and society, resulting in substantial economic, mental, and societal costs (Ferrari et al., 2016; Millier et al., 2014). SZ and BD are believed to be neurodevelopmental disorders influenced by both genetic and environmental factors (Thorup et al., 2015). Identifying early risk markers for these diseases can enhance our understanding of their progression and lay the groundwork for primary preventive interventions.

SZ and BD typically manifest in late teenage years or early 20s, while children at familial high risk may exhibit symptoms even earlier, often before the age of 12 (Robinson and Bergen, 2021; Thorup et al., 2015). Having a family history of BD or SZ is the strongest risk factor for developing these disorders and, according to a meta-analysis, approximately 55% of children at famil-

ial high risk will encounter mental illness in early adulthood, with around one-third experiencing severe mental illness (SMI) (Thorup et al., 2018).

The Danish High Risk and Resilience Study (VIA) is a longitudinal cohort study of 520 7-year-old children born to parents with schizophrenia, bipolar disorder, or no mental disorders (Thorup et al., 2015). Its main objectives are to gain insights into the early disease processes of schizophrenia and bipolar disorder, investigate the developmental trajectory of children with familial high risk across various domains (neurocognition, psychopathology, social cognition, motor function) and examine the influence of genetic and environmental factors on the progression of these disorders. The study seeks to explore symptom formation, cognitive impairments, differences in brain structure and activation patterns (Thorup et al., 2018).

According to the VIA7 study results, children born to parents diagnosed with SZ and BD already demonstrate higher rates of psychiatric diagnosis, cognitive deficits (particularly in FHR SZ), and motor difficulties by age 7 (Burton et al., 2017). When compared to controls, children with FHR of SZ show persistent developmental deficits in manual dexterity and balance. While no observable motor development differences are found among children with FHR of BD as a group, children with definite motor problems across all groups had a higher likelihood of experiencing psychosis, suggesting a connection between childhood motor impairment and neurodevelopmental susceptibility to psychosis (Burton et al., 2023). Studying structural brain changes related to these impairments during early disease formation could provide critical information on differences in neurodevelopment between individuals with and without familial risk as well as their causes.

The central sulcus (CS) is an important landmark for examining structural brain differences in individuals with motor and sensory deficits. It is a prominent anatomical feature of the brain that separates the frontal lobe from the parietal lobe and is symmetrically located in both hemispheres of the brain. It is one of the major sulci (grooves) found in the cerebral cortex. Research has shown that alterations in the shape and size of the central sulcus, which separates the primary motor and somatosensory areas, can impact fine motor control and sensory processing in individuals (Jensen, 2016). Therefore, analysis of the shape and morphology of the CS can contribute to a better understanding of the observed neurodevelopmental abnormalities in the FHR group.

The first step in CS analysis is its detection and segmentation, commonly based on structural magnetic resonance (MR) images. Although the central sulcus is one of the most stable and prominent folds of the human brain, its size and shape vary substantially across individuals and between hemispheres (Caulo et al., 2007). For example, one of the most prominent sections of the central sulcus is the so-called hand knob region, which has significant anatomical variations illustrated in Figure 1.

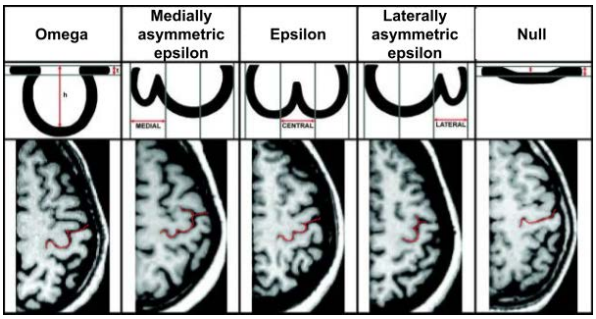


Figure 1: Schematic representation of the different morphological variants of the hand motor cortex observed in humans. Omega, medially asymmetric epsilon, laterally asymmetric epsilon, and null variants were observed in 88.3%, 2.9%, 7.0%, and 1.8% of the hemispheres, respectively with statistically significant sex differences. The epsilon variant was twice as frequent in men, and an interhemispheric concordance for morphologic variants was observed only for women. Courtesy of (Caulo et al., 2007).

Furthermore, the CS morphology depends highly on the gyrification of the cortex, which measures the degree of cortical folding (White et al., 2010) . Increased gyrification characterized by numerous and complex gyri and sulci may lead to more intricate and convoluted sulci patterns with more twists and turns while decreased gyrification, observed with ageing may result in shallower and less complex gyri and wider sulci (Lin et al., 2021b). This decrease in gyrification is caused by systematic cortical thinning during normal ageing and is related to neuronal pruning, life-long reshaping and neurodegenerative processes (Lin et al., 2021b). This in fact means that the intricate pattern of gyri and sulci will vary considerably among different age groups, particularly in children and adults as we know that the peak of gyrification happens in early childhood after which it steadily decreases over time (Klein et al., 2014).

1.2. Project proposal

Segmenting the central sulcus poses a significant challenge due to its intrinsically high morphological variability, which is further influenced by the gyrification changes that occur with age. Successfully addressing these challenges requires sophisticated models and, crucially, large and diverse datasets that encompass the full range of CS morphological variations. Unfortunately, the only currently available dataset with manual sulci segmentations, to the best of our knowledge, is limited in terms of subject count and represents a specific cohort, making robust and precise CS segmentation on diverse populations a difficult task (Brainvisa, 2019).

In light of these challenges, the primary objective of this research is to develop and investigate approaches for constructing robust and adaptable CS segmentation models. Our experiments aim to address the issue of

limited data availability and provide pipelines that can train CS segmentation models that demonstrate reliable performance on unseen and diverse populations of subjects. To achieve them, we investigate two novel ideas in the field of CS segmentation, namely how synthetic data generation can be used to model morphological variability of the CS while self-supervised pre-training and multi-task learning can be utilized for adjusting the model to new subject cohorts by learning pertinent feature representations of the cortex shape.

2. State of the art

Since the development of high-resolution brain MR imaging, the recognition of cerebral sulci and their morphology analysis has been of significant interest to researchers studying structural abnormality patterns related to the diseases affecting the neocortex (Huntgeburth and Petrides, 2012; Mangin et al., 1995). This led to the development of several classes of approaches for automatic sulci detection.

The first type of approaches relies on feature-based elastic registration of a labelled template atlas with segmented sulci to the subject's imaging data. This method propagates labels and identifies anatomical structures of interest by matching surface features between the subject and the pre-labelled template (Behnke et al., 2003; Desikan et al., 2006). While these approaches have been successful in identifying some major sulci, the high inter-subject variability of the cortical folding patterns makes it challenging to achieve an exact match between a subject and a template. Moreover, the existence of such a match is uncertain which further complicates the use of these methods for a precise sulci shape analysis. (Yang and Kruggel, 2008).

Another set of approaches explored by Kao et al. (2007); Shi et al. (2007); Vivodtzev et al. (2003) consider curvature and geodesic depth properties of the cerebral folds. They use depth thresholding and deformable models to differentiate sulci and gyri using cortical surface meshes created from 3D MR images, relying on the assumption that sulci are concave and gyri are convex. However, these approaches highly depend on the ad-hoc handcrafted rules, thresholds and parameters describing the elasticity of the deformable model or depth and curvature thresholds as well as the quality of meshing, which can limit their generalizability and performance.

Recent advancements in image processing, computational methods, and deep learning approaches have led to substantial progress in automatic cortical sulci segmentation (Borne et al., 2020). These advancements increased the accuracy of segmentation as well as expanded on the types and variety of supported sulci, enabling more precise investigations into complex folding patterns and their relationship with brain structure and function (Lyu et al., 2021). In this section, we provide an overview of recent developments in the field, which encompass the most popular pipelines for automatic sulci segmentation and outline the motivations behind the methods explored in this study.

2.1. Spherical CNNs

In the past decade, deep learning models have gained significant traction in biomedical research due to their exceptional ability for feature extraction and outstanding performance (Liu et al., 2021). While there have been previous efforts to apply traditional convolutional neural networks (CNNs) to segment the sulci, such as demonstrated in Yang et al. (2019), the unique characteristics of the convoluted cerebral cortex have led to the proposal to use a spherical variant of CNNs (Lyu et al., 2021).

Standard 2D or 3D CNN architectures are ill-suited for handling the curved geometry of the convoluted cerebral cortex. Most CNN models are designed to optimally work in Euclidean image grids, which restricts their ability to effectively encode cortical surface data. Due to the intricate shape and high curvature of the cortex, it is possible for two points situated on the cortex to have a small Euclidean distance. However, in terms of the manifold distance through the cerebral cortex, these points could be significantly far apart representing distinct and separate regions of the brain. The complex geometry of the cortex introduces a non-linear mapping between Euclidean and manifold distances, meaning that proximity in Euclidean space does not necessarily imply proximity on the cortical surface.

These limitations have prompted the increasing popularity of spherical CNNs as they offer a more suitable framework for processing and analyzing the cortical surface (Willbrand et al., 2022) . However, for them to work, the cortical surface first needs to be represented as a 2D spherical manifold. This process typically involves segmenting the white matter (WM) and grey matter (GM) tissues based on structural brain images, constructing a cerebral cortex surface mesh through tesselation, and applying post-processing steps to address topological inconsistencies, holes, gaps, and optimize surface geometry (McConnell, 1995). Finally, the surface mesh is inflated while preserving its metric properties, resulting in an expanded, spherical representation of the cortex (Fischl et al., 1999).

Lyu et al. (2021) further improves the performance of spherical CNNs in sulci segmentation tasks by applying surface data augmentation and context-aware training in a pipeline schematically depicted in Figure 2. Given the small size of the dataset used (60 and 36 in two explored cohorts), the authors emphasized a crucial need for data augmentation. However, the augmentation approaches for spherical surfaces have not been extensively explored compared to regular 2D/3D data. The authors proposed a novel approach that utilizes surface

registration to augment training samples. The augmentations are achieved by applying spherical harmonics to decompose the spherical deformation needed to register every training image to all others and reconstruct intermediate deformations by controlling the basis functions. By doing so, the suggested approach bridges the gap between moving and target samples in the feature space along their deformation trajectory. This method enhances the training data by generating additional variations that improve the performance of models trained on limited samples. In their context-aware learning phase, hierarchical training is employed. The model is first trained to recognize the deeper and more stable primary sulci, and then the predicted information about their location is used as an additional input channel to guide the segmentation of shallower and more variable tertiary sulci.

While the use of spherical CNNs to capture cerebral surface topology is a promising idea, the numerous pre- and post-processing steps required to segment the tissues and generate cortical meshes and spherical surfaces present drawbacks. The performance of the separate models used in these steps can significantly impact the resulting surface representations of the cortex, leading to missed or wrongly detected sulci regions. The data augmentation technique proposed by the authors although presents a novel augmentation scheme for spherical data is nevertheless limited in its variability to sulci patterns presented in the training data. The limited amount of data augmentation techniques for spherical surfaces and the general lack of research in the field of spherical CNNs can impede the development of robust segmentation algorithms.

2.2. Brainvisa

The BrainVISA software package is widely recognized and utilized in the literature for sulci segmentation (Kochunov et al., 2011; Leroy et al., 2015; Ochiai et al., 2004; Perrot et al., 2011; Roell et al., 2021; Zhang et al., 2020). It offers the capability to segment more than 120 different sulci of the brain and compute morphological features based on the segmentation. In its latest version, as described in Borne et al. (2020), BrainVISA introduces several approaches for sulci labelling, consolidating decades of research in developing automatic pipelines for sulci segmentation. Although these approaches follow different directions for segmentation, they all share the same data preparation steps and begin with sulci detection. BrainVISA's pipeline encompasses multiple pre-processing steps and as shown in Figure 3, as it starts from a high-quality structural T1-weighted image used to first detect and then label the sulci.

2.2.1. Pre-processing

The pre-processing steps applied by BrainVISA aim to transform the structural MR image into a binary CSF skeleton image, where non-zero voxels define the skeleton of the CSF that corresponds to the detected sulci (Borne et al., 2020). To achieve this, several key steps are carried out. The pre-processing pipeline starts with bias field correction to mitigate low-frequency intensity variations in the MRI image. Afterwards, brain and cerebellum identification is performed, followed by the removal of non-brain tissues using a technique based on 3D erosion and template-based 3D region growth. The cortical grey matter ribbon is then obtained after which spherical meshes for the pial and GM/WM interfaces are extracted. Next, based on curvature estimation a crevasse detector reconstructs sulcal structures as medial surfaces between the two opposing gyral banks spanning from the most internal point of the sulcal fold to the cortex's convex hull. Following that, the skeleton of the CSF is fragmented into elementary folds, ensuring adherence to topological and geometric constraints specific to the sulci definition (Zhang et al., 2020). Finally, the CSF skeleton image is parcellated into distinct sulci using one of the following methods.

2.2.2. Multi-atlas parcellation

Multi-atlas segmentation (MAS) methods, originally presented by Rohlfing et al. (2004), leverage manually segmented images as atlases, wherein each atlas is adjusted to fit the image being segmented, and the best matches are selected to participate in the segmentation process. This approach enables a more accurate representation of anatomical variability by avoiding the use of an average template atlas to model the segmentation problem. Instead, MAS techniques incorporate atlases that better capture the inter-subject variability present in the data. It is worth noting, however, that the registration of atlases to the target images can be computationally demanding.

In BrainVISA, the MAS technique involves creating patches extracted as cubical slices from the training images that encompass the elementary sulci detected in the preceding steps (Borne et al., 2020). These patches are then registered to the target image, where the folds skeleton has been extracted, and the best matches are determined. The patch labels are subsequently propagated onto the target image, utilizing the distance between patches to perform a robust weighted average of the labels. Finally, the propagated labels are utilized to calculate the label score maps.

2.2.3. CNN parcellation

Similarly to approaches based on spherical CNNs, BrainVISA's deep learning models do not rely on the original intensity image but instead utilize a preprocessed version of it. They employ a binary 3D image that represents the skeleton of CSF. BrainVISA authors experimented with a 3D U-Net convolutional neural network (CNN) based on the architecture proposed by Çiçek et al. (2016), examining both patch-based and

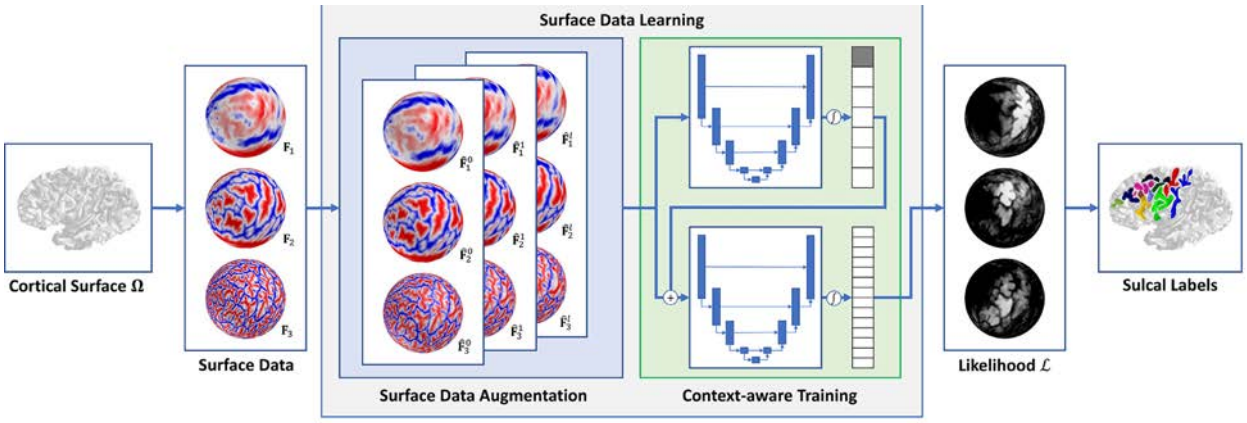


Figure 2: A schematic representation of the framework proposed by Lyu et al. (2021) for the training of spherical CNNs for sulci segmentation. Two main contributions are the data augmentation approach (blue box), which augments training samples by deforming them through surface registration to every possible pair of other training samples while reconstructing all intermediate deformations and using them as additional samples and the context-aware training method (green box) in which spatial information of primary/secondary sulci is extrapolated to guide the segmentation of smaller and shallower tertiary sulci. Courtesy of (Lyu et al., 2021)

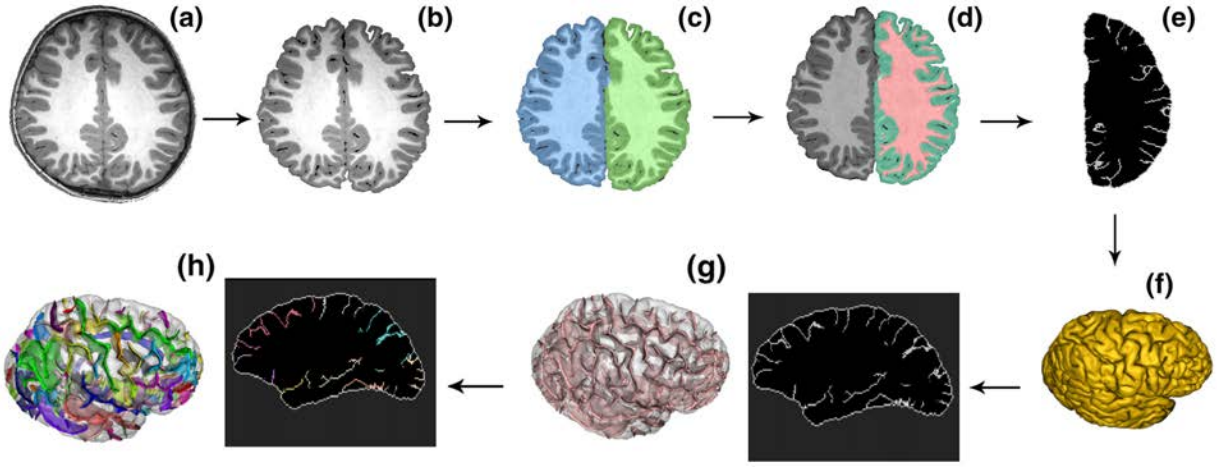


Figure 3: BrainVISA pre-processing pipeline. (a) T1w structural image; (b) Skull stripping; (c) Hemisphere segmentation; (d) GM and WM segmentation; (e) CSF skeleton labelling; (f) Cerebral cortex surface reconstruction; (g) Sulci detection; (h) Sulci parcellation. Based on (Zhang et al., 2020).

whole-image-based models, concluding that the U-Net model processing the entire image outperformed the patch-based one. The superiority of the whole-image approach was attributed to its ability to capture comprehensive sulcal patterns more efficiently. This was achieved by having a larger field of view and the capability to observe the complete folding pattern of the brain, enabling a better understanding of the overall structure (Borne et al., 2020).

During the training process, the CSF skeleton image was used as input to the DL model, which was trained to produce a parcellation of the skeleton by assigning a specific sulci label to each non-zero voxel. During training, only the classification error of the voxels belonging to the skeleton contributed to the loss, based on the assumption that the sulci detection step was executed accurately. Such design choice reduced the complexity of the learning task, as the model solely had to learn the labelling of the skeleton voxels without considering the background. Moreover, due to the heavy reliance on pre-processed skeleton images, the researchers employed only a simple random rotation-based augmentation during training, since the binary nature of the images limited the application of complex data augmentation techniques.

2.3. Limitations of the current methods

While the methods utilizing spherical CNNs and BrainVISA pipelines for automated sulci segmentation demonstrate significant advancements in the field, they are not without their limitations.

First of all, both of them have significant pre- and post-processing pipelines. The performance of individual models employed in them can substantially impact the resulting spherical surface or CSF skeleton cortex representations. Both of them heavily rely on the quality of the WM/GM/CSF segmentations that are used to build cortex meshes which could be a complicated task, especially in low-resolution images corrupted by artefacts, introducing potential inaccuracies or errors. Furthermore, in a population of children or adolescents, for example, higher cortical gyrification can lead to narrower sulci gaps which make proper differentiation between opposing gyral banks challenging due to the partial volume effect (Kochunov et al., 2005).

Secondly, both methods employ a narrow range of data augmentations, which has limited effectiveness in enhancing the diversity of cortex morphologies represented in the training set. These augmentations might fail to adequately simulate the variability of sulci that could be absent in the original data or image variation and bias induced by differences in acquisition schemes or scanners.

Finally, both approaches are trained and evaluated on small-scale in-house datasets consisting of only a few dozen images, typically representing a specific cohort. This limitation arises from the lack of comprehensive, diverse, and standardized datasets available for evaluating sulci segmentation techniques. Consequently, the generalizability and robustness of these approaches across different cohorts are called into question.

These challenges and limitations motivate us to investigate alternative approaches in this study. Our focus is to develope models that could effectively handle variations in image quality and contrast, which would simplify the segmentation pipeline avoiding multiple preprocessing steps that can lead to the accumulation of errors. We are interested in exploring approaches that can efficiently utilise little available data as well as adapt the model to diverse cohorts with previously unseen sulcal patterns. In the subsequent section, we detail the specific methodologies employed to achieve these objectives.

3. Material and methods

3.1. Datasets

In this section, we discuss the datasets used for training and evaluation. It is important to note that the primary objective of this study is to investigate the training of robust and generalizable segmentation models. Therefore, we are exclusively using the BrainVISA dataset that has high-quality curated and labelled CS segmentations to train or fine-tune models for the CS segmentations task, while the VIA11 dataset is used solely for evaluation or self-supervised pre-training that assumes that the CS ground truths do not exist. Such a split allows us to assess the performance degradation of models trained on one dataset and evaluated on another, analysing how inherent disparities in population demographics and acquisition parameters between the datasets affect the model's performance.

To ensure uniformity in the input data for the models, we apply the same pre-processing steps for both datasets, which include only skull-stripping and registration to the common MNI template (Collins et al., 1994). Furthermore, the images were cropped to content and resampled to a consistent resolution of 256x256x124 using the Python implementation of SimpleITK by Yaniv et al. (2017), thereby ensuring identical embedding dimensionality for VIA11 and BrainVISA images.

3.1.1. BrainVISA

Along with presenting the latest sulci segmentation approaches of BrainVISA, Borne et al. (2020) have also released the dataset used to train them. Although it represents a significant contribution in terms of data availability, providing the first to our knowledge high-quality manually segmented dataset with multiple sulci labels for multiple subjects, it has strong limitations in terms of cohort representation.

The dataset contains images from 62 healthy subjects selected from various databases. The subjects are predominantly right-handed men aged between 25 and 35 years. For each subject, a panel of experts produced segmentations for 63 sulci in the right hemisphere and 64 sulci in the left hemisphere through an iterative process involving consensus-based labelling, where agreement among all experts was required for the final segmentation. Although precise, such a labelling scheme excludes the possibility of estimating inter-rater reliability. The dataset includes skull-stripped T1-weighted images, CSF skeleton images, sulci segmentations and brain masks for each subject.

The dataset was randomly split into the train (70%) and validation (30%) sets, enabling performance evaluation on the BrainVISA data as well. Only the training portion of the dataset was used for CS segmentation learning, synthetic data generation and self-supervised pre-training as described in the following sections.

3.1.2. VIA11

The VIA11 study is the second phase of the longitudinal VIA project, which focuses on assessing participants in their 11th year of life (Thorup et al., 2018). In contrast to the initial examination conducted at age 7 (VIA7), VIA11 study protocol incorporates several neuroimaging techniques, with our analysis focusing on structural T1-weighted (MP2RAGE) images.

For our study, we included 303 subjects who participated in the VIA11 study and had structural MR images of sufficient quality. The cohort's average age is 12.1 ± 0.28 years. It has a balanced gender distribution (49% male, 51% female) and includes predominantly right-handed individuals (258 right-handed, 26 left-handed, 19 ambidextrous).

Central sulcus labels for this dataset were obtained using a semi-automatic approach. First, the BrainVISA Morphologist pipeline (Borne et al., 2020) was employed for the initial sulci segmentation of all subjects. Then, manual quality control was performed to estimate their correctness which resulted in 125 subjects having sufficiently good segmentations, 165 subjects having notable errors that would require manual correction of the BrainVISA segmentations, and 13 subjects having incorrect orientation or other errors that prevented manual quality control. In our work, we used 125 subjects' images for which the initial automatic tissue and sulci segmentation procedures were deemed of sufficient quality to perform self-supervised learning as well as estimate the model's performance and compare it among our approaches. The remaining 165 subjects, for which manually corrected segmentations were not available until the very end of the project were never used for any supervised or unsupervised training. These 165 images were only used as a hold-out test set for comparison between BrainVISA and our approaches. It is worth noticing nevertheless, that those manual corrections were performed based on the BrainVISA initial segmentations, and mostly consisted of removing/adding voxels to the BrainVISA's output, which in fact means that these segmentations are highly biased towards BrainVISA's output. We also note that the only pre-processing step applied to VIA11 images was skull-stripping and registration to the MNI space using the BrainVISA software, thereby replicating the same pipeline used to generate the BrainVISA dataset. This ensures uniform data representation for both the training and evaluation phases.

3.2. Methods

Data augmentation has emerged as a popular technique for training deep learning models in scenarios with limited training data, particularly in the medical imaging domain (Chlap et al., 2021). In this section, we explain our rationale for employing synthetic data generation based on the work by Billot et al. (2023a) and provide specific implementation details.

Moreover, to address the issue of limited and constrained diversity in the datasets, we investigate the use of a contrastive self-supervised framework, SimCLR, developed by Chen et al. (2020), in conjunction with our synthetic data to learn cortex representation through self-supervised and multi-task training. We demonstrate how this approach can facilitate model adaptation to new datasets without labelled data, thereby aiding in performing segmentation tasks on dissimilar populations.

Given that our primary focus is exploring training and data generation techniques, we have opted to utilize a simple yet effective 3D CNN U-Net segmentation model designed by Çiçek et al. (2016). 3D U-Nets are among the most commonly employed architectures for 3D medical image segmentation, demonstrating effectiveness, relative computational efficiency, and robustness in the medical domain (Hesamian et al., 2019). U-Net models typically consist of symmetric encoder and decoder parts, that have skipped connections between them. The encoder part of the U-Net is responsible for extracting hierarchical and abstract feature representations from the input image, which is passed to the decoder responsible for upsampling the encoded feature maps to the original input image dimensions and generating the final segmentation map. Specifically, in all our experiments, we employed a 5-level 3D U-Net with an implementation from MONAI, Cardoso et al. (2022), featuring 16, 32, 64, 128, and 256 channels per layer. This choice allowed us to work with a model of comparable size and complexity to that used by Borne et al. (2020), while considering the limitations of our computational resources.

3.2.1. Synthetic data generation

SynthSeg, introduced in the work by Billot et al. (2023a), is a segmentation model that leverages a gen-

erative approach to create synthetic images for network training. By dynamically generating training images with fully randomized parameters, the SynthSeg model learns contrast, intensity, scale, resolution, morphology, artefacts, and noise invariant features, leading to superior segmentation performance, particularly on low-quality images (Billot et al., 2020, 2023b; Iglesias et al., 2021). We adapted the SynthSeg's data generator for our specific problem, utilizing its powerful generation capabilities to create a diverse image dataset based on the limited available labelled images.

Figure 4 illustrates the general pipeline used to create our synthetic dataset. It starts from a segmentation (containing labels of tissues to synthesize, such as WM, GM, CSF, skull bone, and fat) that is passed as input to the SynthSeg data generator. We obtain these segmentations for our datasets from two different sources.

For the VIA11 we utilize FreeSurfer's Samseg tool (Puonti et al., 2016) to obtain preliminary segmentations. These segmentations are then manually quality-checked by a skilled neuroscientist. This quality control ensures that the resulting brain segmentations are anatomically correct and could be used for subsequent image synthesis. From the 125 subjects originally reserved for SST, we select 101 that pass this quality control and only their segmentations are used for the synthetic image generation based on VIA11 data. Billot et al. (2023a) show that with as little as 20 segmentation maps they can reach the top performance therefore we believe that our choice of using only the 101 highest-quality segmentations will not impede the performance of the models. Generated synthetic images from the VIA11 segmentations are then used only for self-supervised pretraining described later.

For the BrainVISA dataset, we obtain the tissue segmentations by utilizing an implementation of an expectation-maximization-based algorithm described by Schindler and Dellaert (2004), that classifies image voxels between WM and GM based on the estimated parameters of the intensities distribution of tissues. Since BrainVISA images contain voxels belonging only to either WM or GM and skull stripping of those images is manually corrected, we opt to use this method for its simplicity and speed. For the BrainVISA segmentations, we additionally combine the central sulcus labels with the tissue labels to create a single segmentation that includes both the tissue information required for generating synthetic images and the sulci labels needed to train the CS segmentation model.

After obtaining the final segmentations we employ SynthSeg's data generator while incorporating the following adjustments:

- We use T1w tissue priors provided by Billot et al. (2023a) to generate images with a contrast similar to T1w by sampling the intensity values for each tissue based on its Gaussian mixture model parameters. Although the original paper demonstrated that the same approach could be used to train a model invariant to any specific contrast by sampling random intensities that do not rely on any priors, we decided to use T1w-based intensities to simplify and speed up the training process, since our goal is evaluating the models' performance on the VIA11 dataset, which also consists of T1w images. Furthermore, we believe that restricting the power of possible augmentations can lead to faster convergence and the ability to learn from fewer data which is favourable in the current setting due to the limited amount of generated images and computational resources

- We preserve the original image dimensions when generating the output, excluding the random resampling and cropping transformations employed by the SynthSeg model. Preserving sufficient spatial resolution is crucial for accurate sulci segmentation, and reducing resolution or cropping the images may result in the loss of important information. Additionally, both the VIA11 and BrainVISA datasets contain isotropic images with a spatial resolution close to 1x1x1mm, eliminating the need to learn resolution-agnostic features for our experiments.

- We utilize the complete set of original spatial transformations, including random affine and elastic transformations of the segmentation map, Gaussian blurring, and bias field corruption applied to the generated image.

- As we use skull-stripped images, no transformations related to random drops of segmentation labels related to the skull are performed.

- Sulci labels are not considered during the image generation step. The model uses corresponding sulci voxels labels of background, WM, or GM for synthesizing intensities under the sulci labels, ensuring the integrity of the image and allowing potential overlap of sulci labels with WM or GM voxels, if present in the original images.

After the generation process, we obtain a pair of synthetic intensity images and the corresponding segmentation, which includes labels for the tissues and the sulci (only for BrainVISA). This dataset obtained through the generative model is referred to as the synthetic dataset.

By applying rigid and non-rigid spatial transformations, random intensity sampling, artefact generation, bias field corruption, and blurring, we simulate high variability in image appearances as well as cortex morphology while preserving crucial information necessary for CS detection and segmentation. Clarisse et al.

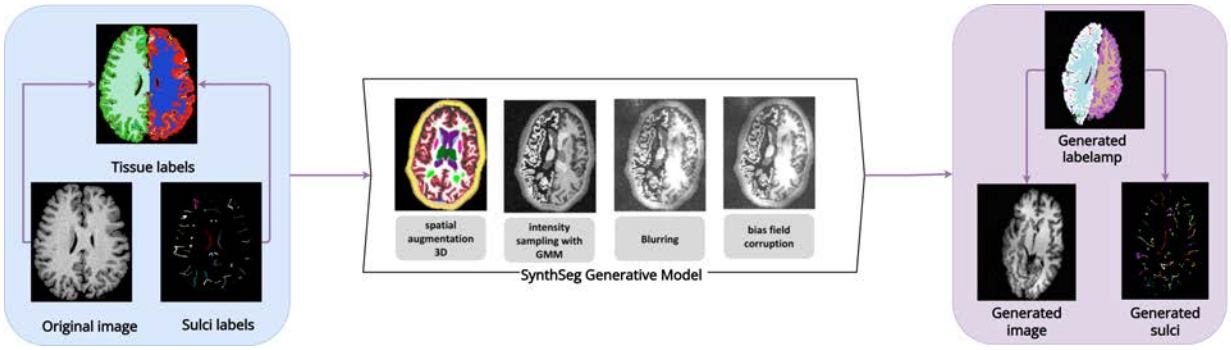


Figure 4: Synthetic Data Generation Pipeline. First, we create a segmentation map, that contains both the tissue and sulci labels. Then we pass it through the SynthSeg generative model, which applies a series of transformations to the segmentation and creates the artificial image by sampling tissue-specific intensity values based on the tissue priors. Finally, the output of the model is the synthetic image and transformed segmentations that contains sulci and tissue labels.

(1997) demonstrated that consistent and accurate identification of the CS relies on several key criteria, including its relative location to other stable and distinct folds, specific shape patterns, as well as its symmetrical location and position on the cortical surface, all of which remain relatively invariant with our transformations. Therefore, by distorting the images in ways that maintain these criteria, we hypothesize that the model will learn a more robust representation of the CS location and shape, invariant to potential distortions that can occur in different datasets, caused by the changes induced by gyrification or brain volume differences, leading to better recognition performance and increased robustness on the morphologically diverse datasets. This hypothesis is supported by the findings of Billot et al. (2023b), who showed the effectiveness of such synthetic data generation for tissue segmentation tasks.

3.2.2. SimCLR

Self-supervised learning (SSL) is a popular method for training DL models in the absence of labelled data that has been especially popular in the medical imaging field, where the cost of labelling is extremely high (Huang et al., 2023). Contrastive training is one of the popular SSL approaches used to learn meaningful representations of input data by maximizing the similarity between different views of the same input and minimizing the similarity between views of different data (Jaiswal et al., 2020). A Simple Framework for Contrastive Learning of Visual Representations (SimCLR) (Chen et al., 2020) is a successful implementation of contrastive learning, particularly in medical image classification and segmentation (Azizi et al., 2021; Dominic et al., 2023; Zeng et al., 2021). The general structure of SimCLR is illustrated in Figure 5. Our objective in utilizing SSL and SimCLR is to integrate knowledge about cortex morphology, including sulci position and shape, into the model weights during the pre-training phase. This integration is expected to be beneficial during the subsequent fine-tuning phase, where the model will focus on learning central sulcus segmentation.

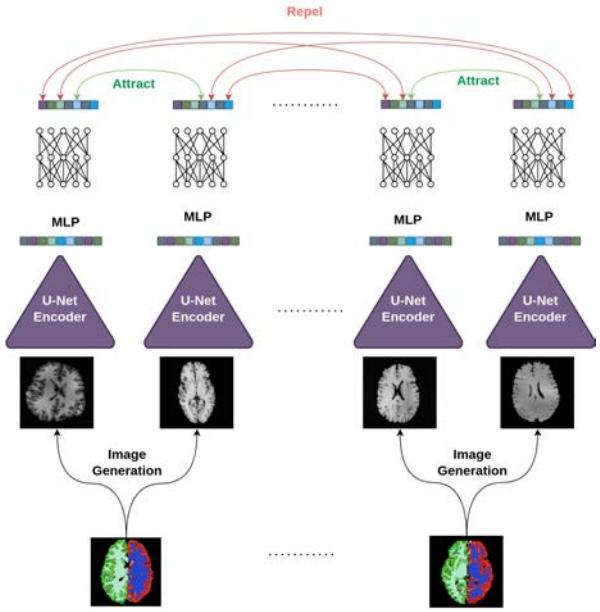


Figure 5: SimCLR framework architecture. First, two image views are generated for each segmentation present in the batch using a synthetic data generator. These synthetic images are then passed through a U-Net encoder, which calculates a dense image representation which is further projected into a space where contrastive loss is computed using an MLP. The loss function encourages the embeddings of images from the same segmentation to be close together in the embedding space while pushing apart the embeddings of images from different segmentation maps.

The first step of the SimCLR framework involves generating multiple views of the same input image, which is a crucial step aimed at preserving relevant semantic information while introducing image variability. Random cropping, colour distortion, and Gaussian blur proved to be effective transformations for generating views in natural image classification (Chen et al., 2020). However, in our case of 3D grayscale volumes in which cropping might erase important information about the cortex morphology we apply a different ap-

proach. Instead, we leverage the synthetic dataset generated from a single tissue segmentation and treat it as a dataset containing multiple views of the same input. We hypothesize that the diverse synthetic images derived from the same segmentation capture essential and identical information about the same cortex morphology while the unique transformations applied to each image introduce the necessary variability. This view generation process can be thought of as creating multiple distortions of the same cortical morphological fingerprint by stretching, scaling, changing its colour or elastically deforming it that would nonetheless preserve the unique pattern present in it.

After generating the different image views, they are sequentially passed through the base model and a non-linear transformation unit based on a Multi-Layer Perceptron (MLP). The base model serves as a robust feature extractor, producing a dense representation of the image that captures key features. Inspired by recent experiments (Dominic et al., 2023; Zeng et al., 2021), we choose the U-Net encoder as the base model. We utilize the first five layers from the downsampling path of the U-Net and flatten the output of the last down convolution layer after max pooling to introduce it as input to the MLP. The MLP projects the feature embeddings from the base model space into a space where contrastive loss is calculated. This helps filter out specific features preferred by the contrastive loss optimization and allows the base model to learn a more robust image representation. In our experiments, we used a 3-layer MLP with a final embedding dimension of 128, as deeper MLPs have shown better results (Azizi et al., 2021).

The final step involves calculating the embeddings' similarity and optimizing the total contrastive loss shown in Equation 1, which is based on the Normalized Temperature-scaled Cross Entropy Loss (NT-Xent) derived from the work of Oord et al. (2018) and displayed in Equation 2. The SimCLR framework aims to maximize the similarity between the embeddings of two augmented versions of the same image (i.e., z_i and z_j) while minimizing it between views of different images z_k. The similarity between embeddings is estimated using cosine similarity defined in Equation 3.

$$\mathcal{L}_{\text{contrastive}} = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k)) + \ell(2k, 2k-1)] \tag{1}$$

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)} \tag{2}$$

$$\text{sim}(z_i, z_j) = \frac{z_i^\top \cdot z_j}{\|z_i\| \cdot \|z_j\|} \tag{3}$$

By optimizing this loss function, we aim to pre-train the model on a task similar to the downstream task, but without actual labels. The pre-training with SimCLR on synthetic data encourages the U-Net encoder to learn robust and comprehensive feature representations of the cortex morphology, as it is the only consistent and distinctive feature across different image views. Following pre-training, the model undergoes fine-tuning on the downstream task.

Our hypothesis is that by initializing the weights of the U-Net encoder with those learned during pre-training, we can transfer information about the anatomical variability of the cortical folds from a bigger and more diverse dataset and then leverage that knowledge for segmentation through fine-tuning. Specifically, we focus on pre-training with the VIA11 dataset and subsequent fine-tuning on the BrainVISA dataset to assess if the model can capture cohort-specific sulci properties that may be absent in the limited labelled data as we are especially interested in improving the performance on the VIA11 that we use for the final evaluation.

3.2.3. Multi-task learning

In our previous approach, we utilized the SimCLR framework for pre-training the U-Net encoder. However, we also investigate the pre-training of the decoder component in the U-Net architecture. Drawing inspiration from recent advancements in multi-task learning (Gao et al., 2020; Zhou et al., 2021), we propose a novel pre-training framework that combines contrastive self-supervised learning with segmentation learning to pre-train the entire U-Net model.

Illustrated in Figure 6, our multi-task SSL pipeline consists of two parts. The first part follows the contrastive pre-training structure described before, calculating the contrastive loss and updating the weights of the U-Net encoder. The second part employs the same encoder model, combined with a symmetrical decoder which is simultaneously trained in a joint optimization procedure for brain tissue segmentation. We utilize the same labels used to create synthetic images to train the U-Net decoder to segment GM tissue based on the intensity images, effectively replicating a part of the SynthSeg training pipeline. We choose to train the model for only single-class GM segmentation to make it compatible with the downstream single-class task of CS segmentation, avoiding the need to adjust internal embedding and kernel dimensions. Furthermore, GM segmentations were already available as a prerequisite for synthetic data generation and segmenting GM requires an understanding of the cortex morphology from the model at the surface detection level, albeit based on intensity contrast.

The final loss function optimized in this pipeline, shown in Equation 4, is a combination of the segmentation loss and the contrastive loss discussed earlier. We employ the soft dice loss implementation from MONAI (Cardoso et al., 2022) for the segmentation loss. This training scheme enables us to target the full U-Net

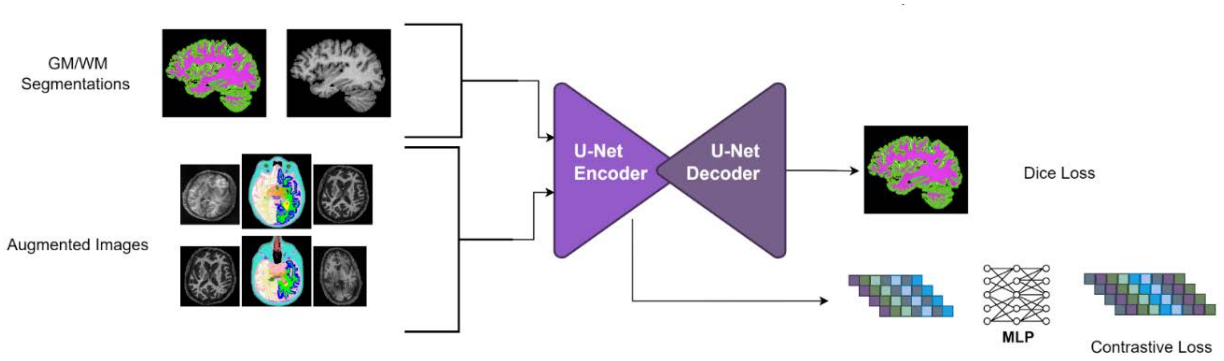


Figure 6: Multi-task self-supervised training scheme. Combined contrastive and segmentation loss allows pre-training of both the encoder (on contrastive and GM segmentation tasks) and decoder (only on GM segmentation task) of the U-Net.

model during the pre-training phase, learning improved weights initialization for both the encoder and decoder. Moreover, it encourages the encoder to learn representative features not only for the contrastive task but also for the segmentation task directly. By employing the multi-task loss, we aim to leverage the information learned during the pre-training phase more effectively in the downstream phase.

$$\mathcal{L}_{\text{multi-task}} = \mathcal{L}_{\text{segmentation}} + \mathcal{L}_{\text{contrastive}} \tag{4}$$

3.3. Training and Validation Strategy

This section presents the technical implementation details of the tested models, training and validation strategies employed. These details aim to clarify the rationale behind our parameter choices and facilitate the replication of our results.

Unlike the approach proposed by Billot et al. (2023a), which utilizes online data generation which creates synthetic images on the fly and directly feeds them into the segmentation model, we employ an offline generation approach. This choice is driven by computational limitations that prevent us from simultaneously running both the generative and segmentation models on the same GPU. Due to the storage constraints, we generate 100 synthetic images for each subject from both the BrainVISA and VIA11 datasets, resulting in 6,200 synthetic images for BrainVISA and 10,100 synthetic images for VIA11 datasets.

To ensure consistency in training parameters, we train all U-Net models for a maximum of 200 epochs during the central sulcus (CS) segmentation learning process. We employ an early stopping criterion, wherein training is halted if the validation loss fails to improve for the last 10 epochs. For CS segmentation, we adopt the Tversky Loss introduced by Salehi et al. (2017) as our learning criterion. This loss function has demonstrated superior performance in highly imbalanced segmentation problems, which is important in our case as CS voxels occupy on average around 0.02% of all image voxels. Although we train the model on both synthetic and original images in some experiments, validation is always performed using the original images from the corresponding validation splits. We employ a batch size of 1, as it is the maximum that can fit within our available GPUs (NVIDIA GeForce RTX 3090 with 24GB of video RAM) and an initial learning rate (LR) of 0.001.

In the final evaluation stage, we incorporate a post-processing step in our workflow to facilitate a meaningful comparison between the segmentations and meshes generated by our models and BrainVISA's pipeline. Given that our segmentations do not rely on CSF skeleton images and do not impose anatomical correctness requirements as part of their design, it is essential to ensure their compatibility with the meshing algorithm. To achieve this, we have opted to employ the same meshing algorithm utilized by BrainVISA in their pipeline for generating meshes from segmentations. By adopting the same tool, we minimize additional variability introduced by different meshing techniques, enabling a more accurate comparison of mesh properties.

The meshing process involves intricate calculations to create a surface based on a point cloud (McConnell, 1995). However, during this process, errors such as gaps, holes, and excessive tessellation can arise, particularly in the presence of noise points. To address these issues and achieve appropriate tessellation in our generated segmentations, we apply a straightforward post-processing approach. We begin by performing a morphological binary dilation on the obtained segmentation, connecting sulcus segments that are close to each other but separated spatially. Subsequently, connected component labelling is applied to the dilated image. In the final segmentation, we retain only the voxels from the original segmentation that belong to the two largest connected components calculated from the dilated image. This step ensures that only the central sulcus segments from the left and right hemispheres are retained before the meshing stage, reducing errors in the resulting sulcus mesh and enhancing its quality. We apply this post-processing step only for the final comparison between BrainVISA's pipeline and our approaches as it

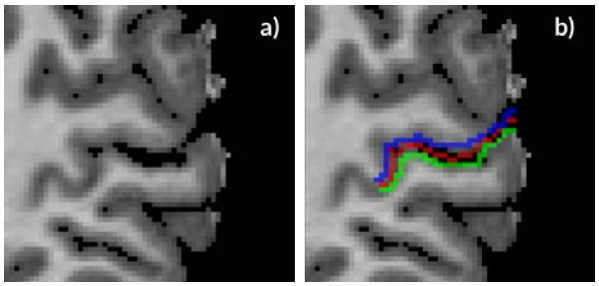


Figure 7: Ambiguity in CS segmentation. a) Brain image; b) Brain image with overlapped sulci segmentations: ground truth (red) and manually drawn alternatives (blue and green). Notice how the alternative segmentations closely follow the ground truth in terms of shape and correct anatomical position, despite having zero overlap with the ground truth and yielding a DSC of 0. The precise localization of the CS ribbon within the sulcal gap, which often spans multiple voxels in width, is inherently ambiguous. Therefore, a metric that accounts for the distance between segmentations provides a more robust measure, which makes it crucial to consider.

is needed specifically for correct meshing and fair comparison with the full BrainVISA pipeline.

3.4. Quantitative Analysis

To evaluate the quality of our segmentations, we employ two widely used metrics: the Dice similarity coefficient (DSC) and the Hausdorff distance (HD).

The DSC quantifies the voxel-wise overlap between two segmentations, denoted as X and Y. Its values range from 0 to 1, where 0 represents no overlap and 1 indicates complete agreement. The DSC is computed using the following formula:

$$DSC = \frac{2 \times |X \cap Y|}{|X| + |Y|}$$

Another important metric we employ is the Hausdorff distance, defined as:

$$H(X, Y) = \max \left\{ \sup_{x \in X} \inf_{y \in Y} \rho(x, y), \sup_{y \in Y} \inf_{x \in X} \rho(x, y) \right\}$$

It measures the mutual proximity of two segmentations and provides insight into their spatial dissimilarity. HD reflects the maximum distance of the two closest points in the segmentations and it takes positive values with smaller ones reflecting higher proximity and 0 corresponding to the complete overlap of two segmentations. Given the nature of the segmentation task, we argue that the Hausdorff distance is a crucial measure of segmentation quality that should be considered. Sulci localization is a complicated task and the precise placement of the sulci ribbon in the gap between two gyri is often ambiguous as shown in Figure 7.

3.5. Implementation Details

We used Python programming language with several frameworks for this project. Pytorch and Pytorch Lightning were used to implement and train the SSL and DL models while Tensorflow was used to adapt and run the synthetic data generation pipeline. Additionally, libraries like SimpleITK-SimpleElastix and nibabel were used for image registration and spatial transformations and ITK-Snap with 3D Slicer were used for visualization purposes.

Project code as well as other hyperparameters values and corresponding documentation can be found at: https://github.com/Vivikar/central-sulcus-analysis.

4. Results

In this section, we present the results of our experiments, both qualitatively and quantitatively, following the same order as in the previous section.

4.1. Synthetic data generation

We begin by examining the impact of synthetic data on the model's generalizability. To assess this, we compare the performance of the model trained on the synthetic BrainVISA dataset with the model trained on the original BrainVISA dataset. Figure 8 displays the quantitative results comparing the performance of these two models on two evaluation datasets: one composed of the original BrainVISA images from the validation split and the other consisting of original 125 images from the VIA11 dataset, for which we have used BrainVISA's segmentations as ground truth since they passed the quality control.

On both datasets, we observe a decrease in the Dice similarity coefficient for the models trained on synthetic data. This finding aligns with the studies conducted by Billot et al. (2023a) and Billot et al. (2023b), which demonstrate that while synthetic data yields significant improvements for images affected by artefacts, low quality, or low resolution, models trained on synthetic data tend to under-perform compared to state-of-the-art models on high-quality and high-resolution images. However, the model trained on synthetic data exhibits a substantial decrease in Hausdorff distance scores on the VIA11 dataset, which arguably provides a more sensible evaluation of performance in this setting (see Figure 9).

Figure 9 presents qualitative results that help explain these outcomes. It illustrates how the model trained solely on the original data misclassifies the region unrelated to the central sulcus (CS), while the model trained on synthetic data does not. It is important to note that the misclassified region corresponds to the neck and represents a skull-stripping error, as it should not be present in the skull-stripped image. However, the model trained on synthetic data makes errors in mistakenly segmenting sulci neighbouring to CS as illustrated in image c) of Figure 9. Although these errors lead to significantly lower Hausdorff distance scores as they are closer to ground truth, they are of great concern as they

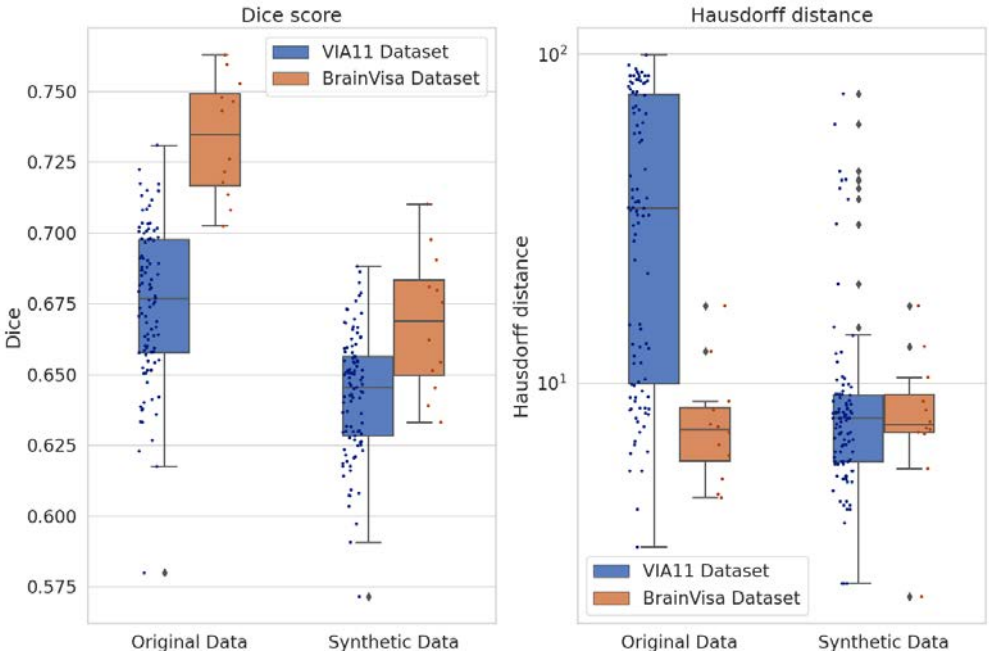


Figure 8: Box plot showing DSC and HD scores for the models trained on the synthetic and original BrainVISA datasets and evaluated on the original images from the BrainVISA validation split and VIA11 dataset. A statistically significant decrease (p-value < 0.000001 based on a two-sided t-test) of HD scores between the model trained on synthetic and original data is observed for the VIA11 dataset.

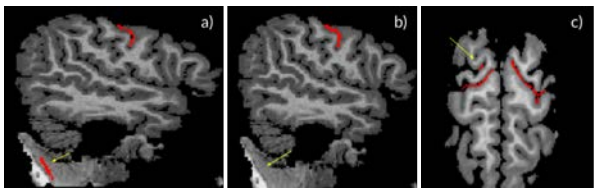


Figure 9: Sample segmentations for VIA11 subjects for the a) model trained on original BrainVISA data, and b) and c) model trained on synthetic BrainVISA data. The yellow arrows indicate the regions where some of the modes made mistakes.

still have a substantial impact on the meshing of the CS segmentation and the subsequent estimation of its morphological features.

In our following experiments, we use the synthetic dataset for learning CS segmentation in the fine-tuning stages of the SSL as it demonstrates superior performance.

4.2. SimCLR

To test how SSL can aid in adjusting the model to new datasets we apply the synthetic data generation approach described earlier to the 101 VIA11 images to create a synthetic VIA11 dataset, which we utilize for self-supervision in conjunction with the synthetic BrainVISA dataset.

4.2.1. SSL pre-training

Chen et al. (2020) use several methods to validate the performance of their self-supervised pre-training. However, since our downstream task is not related to classification and the used images do not represent distinct categories of objects, we have chosen to validate the quality of learned image representations through the dimensionality reduction approach. Figure 10 shows the projection of the embeddings outputted by the MLP from 128D space to 2D space using T-SNE (van der Maaten and Hinton, 2008). We select random four validation VIA images that were not included in the 101 images used for self-supervised training, and we generate 100 synthetic images based on the segmentations of these four. Despite the model never encountering images generated from these four segmentations during training, we observe a clear separation between the projected embeddings corresponding to each segmentation which shows that during SSL the U-Net encoder was able to learn distinct features separating these images.

4.2.2. Full U-Net fine-tuning

Figure 11 presents the metrics of the U-Net models with different SSL approaches, fine-tuned on the synthetic BrainVISA dataset with no frozen layers (i.e., all encoder weights were initialized with those calculated during SSL and then updated during downstream training). We observe a statistically significant increase in the Dice score metrics with SSL, with the best values for the VIA11 dataset obtained when pre-trained on VIA11 compared to no SSL (p-value=0.000799). However, we find no statistically significant difference in HD when comparing No SSL and VIA11 SSL for the VIA11 dataset. It appears that SSL with a small and homogeneous dataset like BrainVISA does not contribute to

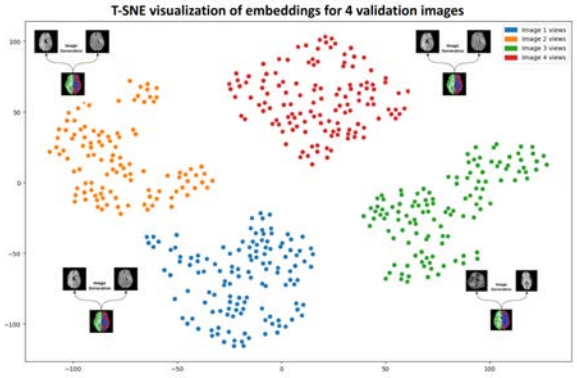


Figure 10: Visualization of embeddings of 100 different images projected onto a 2D plane with T-SNE from 4 different validation segmentations. Each coloured dot represents a synthetic image generated from a distinct segmentation.

the increase of the model's generalizability and robustness, while SSL with a bigger and more diverse VIA11 dataset leads to an increase in the Dice scores without degrading HD. Therefore, in our following experiments, we consider the model pre-trained with VIA11 SSL and fine-tuned with synthetic BrainVISA data as our best-performing model.

4.2.3. U-Net fine-tuning with the frozen encoder

Azizi et al. (2021) highlights the importance of careful fine-tuning when learning downstream tasks to preserve the information learned during self-supervision. We investigate the impact of freezing the encoder model after SSL and thus completely preserving the SSL features for learning the CS segmentation task. Figure 12 compares the models with and without SSL, evaluating whether freezing the encoder during the downstream task benefits the CS segmentation performance. The results show a significant decrease in performance for both datasets and on both evaluation metrics when the encoder is frozen.

4.3. Multi-task SSL

As we can see in Figure 13 there are no statistically significant improvements in any of the metrics for the multi-task learning scenario. Due to computational limitations, we conducted experiments with SSL and evaluation solely on the VIA11 data, focusing primarily on the adaptability of our model to diverse and unseen datasets. Although no improvements were observed, we note that this strategy did not substantially degrade our results therefore such an outcome could be a result of a poor hyper-parameters selection.

4.4. Comparison with BrainVISA

To evaluate the effectiveness of our approach and compare it with the state-of-the-art BrainVISA's pipeline, we conducted several experiments using 165 VIA11 images from the held-out test set. These images were not utilized during any stage of the pre-training or data synthesis. We have obtained manual ground truth CS segmentations by correcting the initial BrainVISA's pipeline output for them, as it was deemed necessary during our initial quality assessment of the BrainVISA's results.

Figure 14 presents a comparison between our best model that is based on VIA11 SST with the further fine-tuning on the synthetic BrainVISA data. We see that BrainVISA's segmentations have a much higher Dice score.

This substantial difference in DSC can be attributed to two main factors. First, the manual segmentations are essentially modifications of BrainVISA results, and in many cases, the initial BrainVISA estimate, if sufficiently accurate, was left unchanged. Second, the nature of the segmentations generated by our algorithm is characterized by thicker segmentation ribbons as can be seen in Figure 15, which are anatomically and morphologically correct but receive lower Dice scores due to its voxel-wise intersection evaluation. However, our approach shows a statistically significant improvement in the HD (BrainVISA's mean 8.315 vs our model's 7.37 with p-value < 0.005).

Considering our ultimate goal of evaluating morphological features of the CS, an essential step in their analysis is meshing and subsequent extraction of shape features. The BrainVISA software provides built-in tools for meshing sulci segmentations. We utilized these tools to compute meshes for the segmentations obtained from the BrainVISA pipeline, manually corrected segmentations, and segmentations produced by our best model. Figure 16 displays correlation plots between the volume and surface area calculated from these three meshes. It is immediately apparent that the morphological features of volume and surface area calculated from our segmentations exhibit a close correlation with those calculated from the manual segmentations.

5. Discussion and conclusions

In this study, we have presented and evaluated various approaches for training deep learning models to perform central sulcus segmentation. Our review of the current state-of-the-art approaches revealed important limitations in models trained on small and restricted labelled datasets, which fail to account for the neuroanatomical variability of cortical morphology. We have emphasized the need for robust and automatic segmentation models and proposed novel frameworks to address this challenge. Our frameworks focus on two key ideas: efficient utilization of limited labelled data through artificial simulation of cortical variability in synthetic images and the creation of a pipeline for adapting the model to new subject populations through self-supervised learning of cortex morphology features.

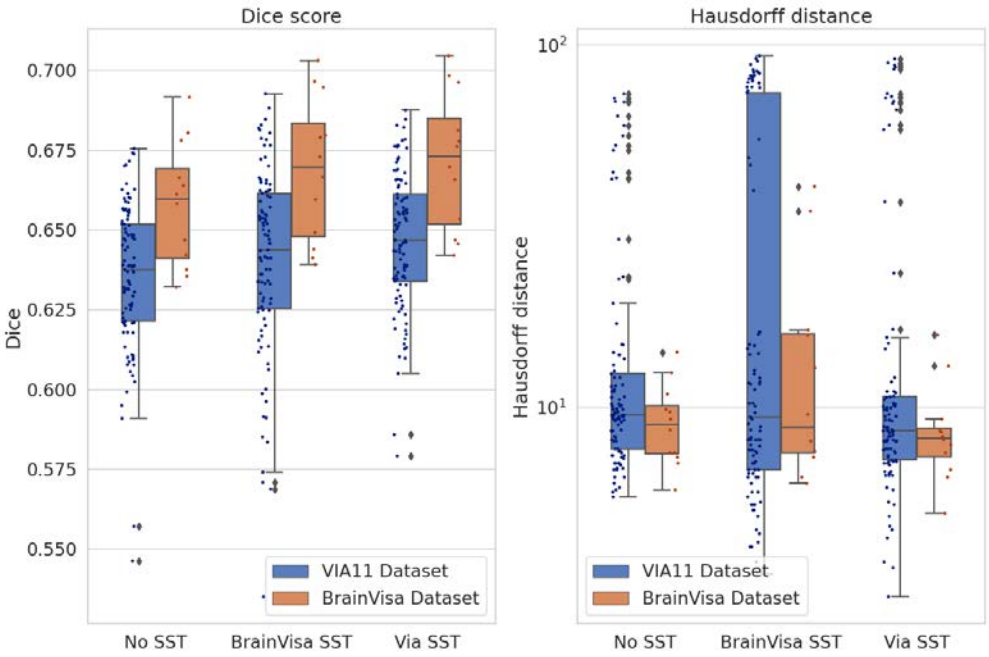


Figure 11: DSC and HD for models with SSL based on different datasets with subsequent fine-tuning of the full encoder and decoder on both original datasets. A statistically significant difference (p-value < 0.005) was observed when comparing No SST with the VIA SST in terms of DSC on the VIA11 dataset.

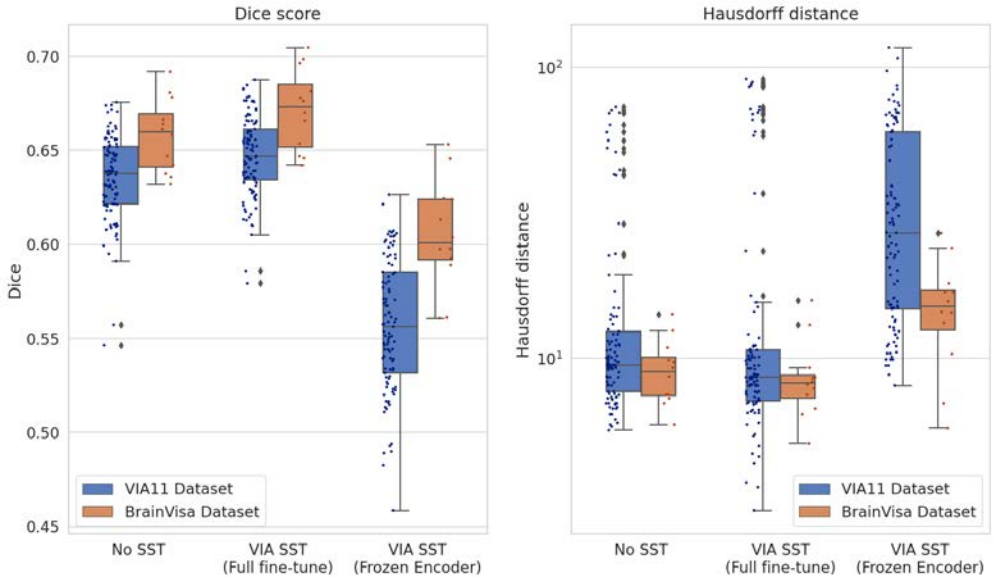


Figure 12: DSC and HD for models without SSL and SSL on the VIA11 dataset with frozen and not frozen encoder on both datasets.

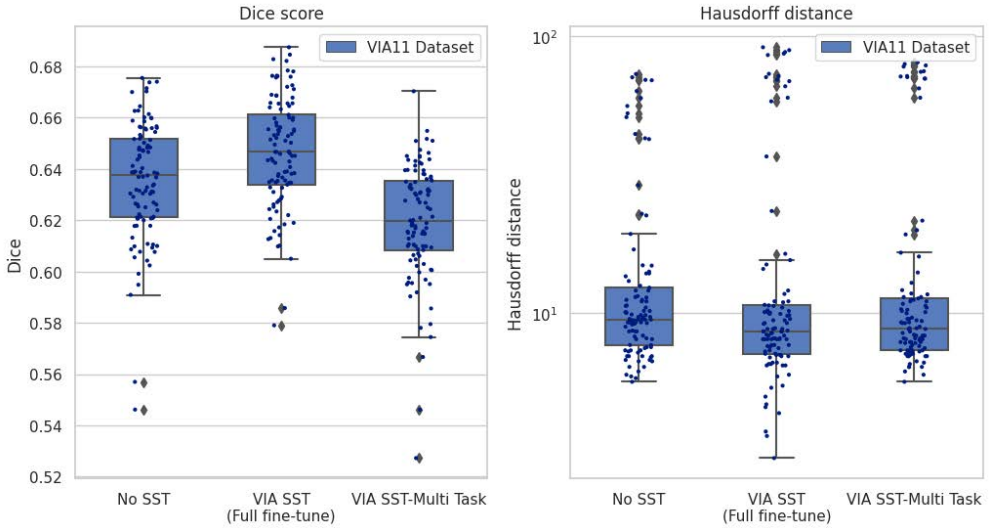


Figure 13: DSC and HD for models trained without SSL, with VIA SSL and with multi-task VIA11 SSL and tested on the VIA11 dataset.

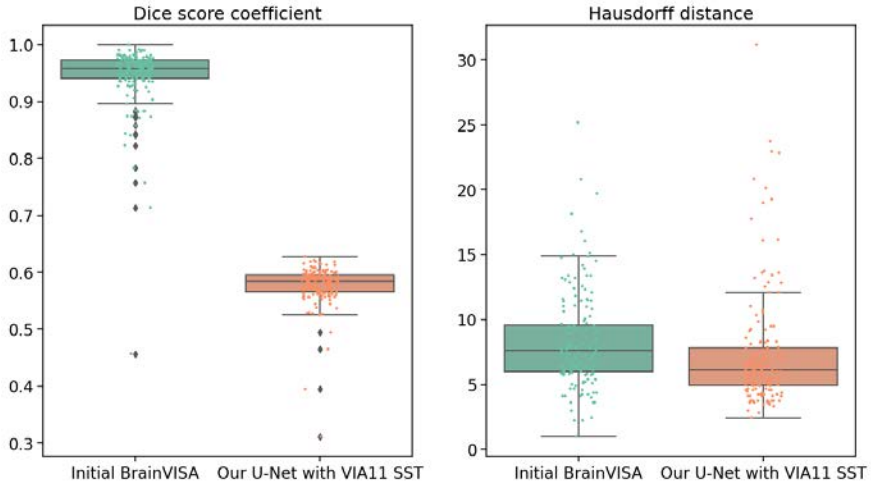


Figure 14: DSC and HD scores of the BrainVISA segmentations and our U-Net model trained with VIA11 SSL. Our model shows a statistically significant decrease in the HD scores with a p-value of 0.0379 based on a two-sample T-test.

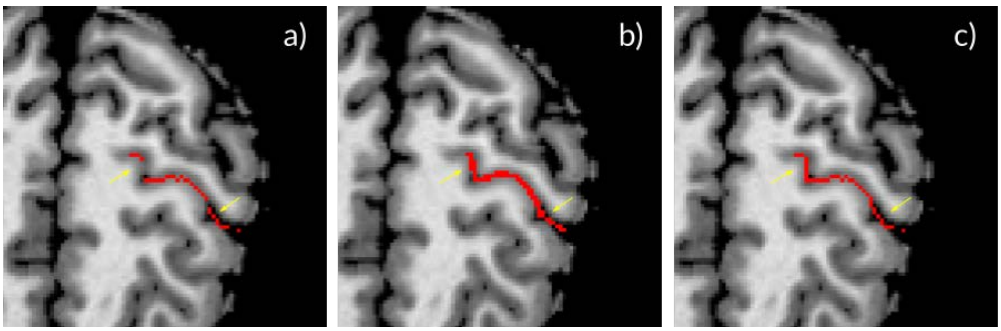


Figure 15: Segmentation examples. a) produced by BrainVISA software, b) produced by our VIA11 SST U-Net model, c) manually corrected ground truth. Yellow arrows indicate gaps in the segmentation ribbon that can lead to holes in the resulting mesh interfering with morphological features calculation.

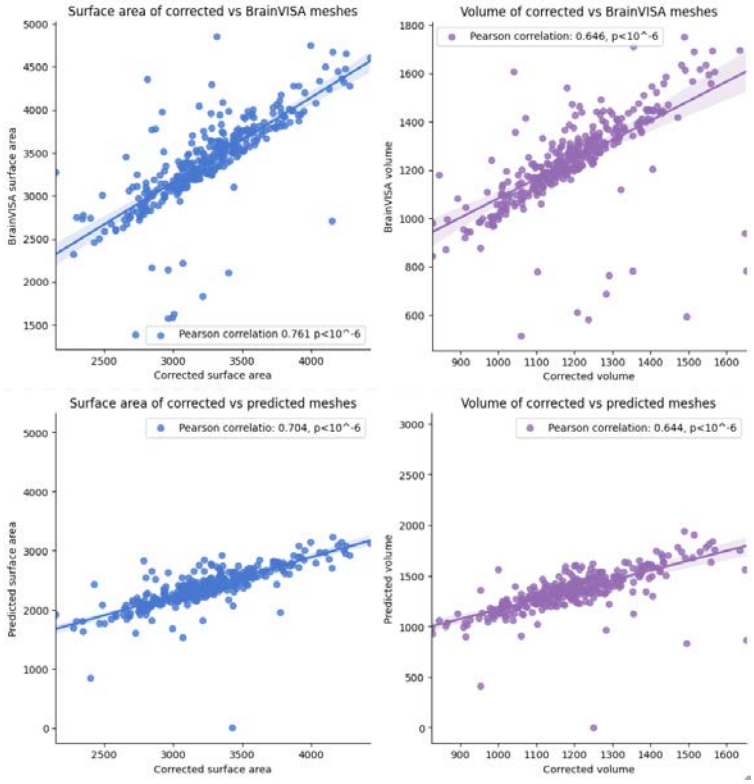


Figure 16: Volume and surface area of the meshes calculated based on the manually corrected segmentation, BrainVISA's and ours (predicted by the VIA11 SST U-Net) plotted against each other.

5.1. Discussion of results

Firstly, we explored the use of synthetic data to train more robust models. Our findings indicate that models trained with synthetic data exhibit significantly lower Hausdorff distance (HD) scores on the VIA11 dataset, despite never being exposed to it during training. This dataset consists of a different subject population, with different image contrasts, and quality compared to the training dataset. This promising result highlights the potential of using synthetic data to simulate the morphological variability of cortex present in diverse subject populations that is beneficial for sulci localization. This approach helps address the limitations posed by small datasets, which have historically hindered progress in sulci segmentation research.

We tested the effectiveness of a self-supervised learning framework based on SimCLR combined with synthetic data for learning unique and distinct representations of cortex morphology. Our experiments demonstrate that our pre-training strategy for the U-Net encoder leads to improved DSC for the VIA11 dataset. This shows that with this pipeline we can adapt our segmentation model to new datasets without requiring any labels for them, effectively transferring information about cortex shape variability from the new dataset to our model. Consequently, this approach shows that we can improve the segmentation results for new cohorts by performing SSL on just the intensity images, paving the way to utilizing abundant unlabelled datasets that are openly available for the training of foundation models that better capture real-world anatomical variability of the cortex.

Although the multi-task SSL approach did not substantially improve our results, we believe that more careful pre-training of both the encoder and decoder models could yield better performance. We did not extensively experiment with hyper-parameter tuning for the multi-task framework and subsequent fine-tuning, which could explain the lack of improvement. Additionally, the substantial difference between GM segmentation and CS segmentation may have hindered the adaptability of the SSL loss for CS segmentation.

Lastly, we compared our best model with the current state-of-the-art pipeline from BrainVISA and demonstrated comparable performance with an improvement in the HD metric. To validate the correctness of the morphological structures represented by these segmentations, we constructed meshes based on them. The meshes built from segmentations produced by our models have highly correlated surface area and volume measures to those obtained from manual ground truths, indicating that the developed approach can be effectively utilized for CS segmentation that can be further used for analyzing the shape properties of the central sulcus.

5.1.1. Limitations and future work

The objective of this study was to explore and establish a proof of concept for training robust CS segmentation models directly from intensity images, without the need for extensive pre- or post-processing steps. We aimed to address the challenges commonly encountered in the medical imaging domain, including the limited availability of labelled data and the high morphological variability of the target structure.

However, it is important to acknowledge the limitations of our work. We conducted evaluations on only one external dataset (VIA11) in addition to the training dataset (BrainVISA). To obtain a more comprehensive understanding of the pipeline's performance and robustness, further evaluations on additional datasets with diverse population cohorts are necessary. For instance, testing the model on datasets consisting of elderly individuals with neurodegenerative processes resulting in substantial brain atrophy or infants and young children with either under-developed or over-developed cortical gyrification could provide valuable insights.

In our synthetic data generation process, we utilized a limited set of artificial images due to storage and computational constraints. To enhance the diversity of the synthetic dataset, implementing an online generation procedure with unlimited and unique images for each generation could be explored. Additionally, we were unable to extensively experiment and determine an optimal set of transformation parameters for the SynthSeg Generative model. We believe that exhaustive tuning of spatial and intensity parameters applied to the images can further increase the dataset's diversity. Moreover, incorporating brain images without skull stripping could improve the model's robustness to potential artefacts that can occur if it is performed with errors as well as potentially eliminate the need for this pre-processing step.

In the SSL training stage, Chen et al. (2020) demonstrated that large batch size leads to better performance. Although longer training can partially mitigate the effects of smaller batch sizes, we did not specifically study how batch size affects our SSL stage. Moreover, in the multi-task SSL setting, we did not explore weighting schemes for the contrastive and segmentation losses due to time and GPU constraints, despite studies suggesting potential benefits (Lin et al., 2021a).

Finally, we have not experimented with different DL architectures for our base segmentation model, although there are many new architectures based on Transformers that show promising results in the medical image segmentation field (Liu et al., 2021) as well as U-Net variations. We have chosen to use a simple and lightweight U-Net model that made possible an extensive exploration of the proposed solutions given our computational constraints.

5.2. Conclusions

Synthetic data generation and self-supervised learning are two powerful tools that can address challenges in the development and deployment of DL models for recognition and segmentation tasks. In this study, we have demonstrated that by employing synthetic data within a self-supervised learning framework that enables the model to learn unique cortical morphology representations, we can achieve results that are comparable to state-of-the-art methods in central sulcus segmentation. These approaches alleviate the need for costly and error-prone pre-processing steps, allowing the training of robust and generalizable DL models that can be adapted to new cohorts without requiring any ground truth labels and work efficiently even with little available training data.

Acknowledgements

First and foremost, I would like to express gratitude to my supervisor, Kristoffer Madsen, for his guidance and invaluable support throughout this project. I would also like to express my thanks to my colleagues at DR-CMR, Enedino Hernández-Torres and Line K. Johnsen, for their assistance with data access, quality estimations, and manual segmentations of the sulci data. I am also grateful to Melissa Larsen and the centre's director Hartwig R. Siebner, for helping me better understand the neurobiological basis of this project as well as for providing me with a great opportunity to work at DR-CMR.

References

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al., 2021. Big self-supervised models advance medical image classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3478–3488.

Behnke, K.J., Rettmann, M.E., Pham, D.L., Shen, D., Resnick, S.M., Davatzikos, C., Prince, J.L., 2003. Automatic classification of sulcal regions of the human brain cortex using pattern recognition, in: Medical imaging 2003: Image processing, SPIE. pp. 1499–1510.

Billot, B., Greve, D., Van Leemput, K., Fischl, B., Iglesias, J.E., Dalca, A.V., 2020. A learning strategy for contrast-agnostic mri segmentation. arXiv preprint arXiv:2003.01995 .

Billot, B., Greve, D.N., Puonti, O., Thielscher, A., Van Leemput, K., Fischl, B., Dalca, A.V., Iglesias, J.E., 2023a. Synthseg: Segmentation of brain MRI scans of any contrast and resolution without retraining. Medical Image Analysis 86, 102789. doi:10.1016/j.media.2023.102789.

Billot, B., Magdamo, C., Cheng, Y., Arnold, S.E., Das, S., Iglesias, J.E., 2023b. Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain mri datasets. Proceedings of the National Academy of Sciences 120, e2216399120. doi:10.1073/pnas.2216399120.

Borne, L., Rivière, D., Mancip, M., Mangin, J.F., 2020. Automatic labeling of cortical sulci using patch- or CNN-based segmentation techniques combined with bottom-up geometric constraints. Medical Image Analysis 62, 101651. doi:10.1016/j.media.2020.101651.

Brainvisa, 2019. Sulci database. Online.

Burton, B.K., Krantz, M.F., Skovgaard, L.T., Brandt, J.M., Gregersen, M., Søndergaard, A., Knudsen, C.B., Andreassen, A.K., Veddum, L., Rohd, S.B., Wilms, M., Tjott, C., Hjorthøj, C., Ohland, J., Greve, A., Hemager, N., Bliksted, V.F., Mors, O., Plessen, K.J., Thorup, A.A.E., Nordentoft, M., 2023. Impaired motor development in children with familial high risk of schizophrenia or bipolar disorder and the association with psychotic experiences: a 4-year danish observational follow-up study. The Lancet Psychiatry 10, 108–118. doi:10.1016/s2215-0366(22)00402-3.

Burton, B.K., Thorup, A.A.E., Jepsen, J.R., Poulsen, G., Ellersgaard, D., Spang, K.S., Christiani, C.J., Hemager, N., Gantriis, D., Greve, A., Mors, O., Nordentoft, M., Plessen, K.J., 2017. Impairments of motor function among children with a familial risk of schizophrenia or bipolar disorder at 7 years old in denmark: an observational cohort study. The Lancet Psychiatry 4, 400–408. doi:10.1016/s2215-0366(17)30103-7.

Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., He, Y., Xu, Z., Hatamizadeh, A., Myronenko, A., Zhu, W., Liu, Y., Zheng, M., Tang, Y., Yang, I., Zephyr, M., Hashemian, B., Alle, S., Darestani, M.Z., Budd, C., Modat, M., Vercauteren, T., Wang, G., Li, Y., Hu, Y., Fu, Y., Gorman, B., Johnson, H., Genereaux, B., Erdal, B.S., Gupta, V., Diaz-Pinto, A., Dourson, A., Maier-Hein, L., Jaeger, P.F., Baumgartner, M., Kalpathy-Cramer, J., Flores, M., Kirby, J., Cooper, L.A.D., Roth, H.R., Xu, D., Bericat, D., Floca, R., Zhou, S.K., Shuaib, H., Farahani, K., Maier-Hein, K.H., Aylward, S., Dogra, P., Ourselin, S., Feng, A., 2022. Monai: An open-source framework for deep learning in healthcare.

Caulo, M., Briganti, C., Mattei, P., Perfetti, B., Ferretti, A., Romani, G., Tartaro, A., Colosimo, C., 2007. New morphologic variants of the hand motor cortex as seen with MR imaging in a large study population. American Journal of Neuroradiology 28, 1480–1485. doi:10.3174/ajnr.a0597.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR. pp. 1597–1607.

Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., Haworth, A., 2021. A review of medical image data augmentation techniques for deep learning applications. Journal of Medical Imaging and Radiation Oncology 65, 545–563. doi:https://doi.org/10.1111/1754-9485.13261.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3d u-net: learning dense volumetric segmentation from sparse annotation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19, Springer. pp. 424–432.

Clarisse, J., Pertuzon, B., Ayachi, M., Francke, J., et al., 1997. Identification of the central sulcus using the scanner and mri. Journal of Neuroradiology= Journal de Neuroradiologie 24, 187–204.

Collins, D.L., Neelin, P., Peters, T.M., Evans, A.C., 1994. Automatic 3d intersubject registration of mr volumetric data in standardized talairach space. Journal of computer assisted tomography 18, 192–205.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage 31, 968–980. doi:10.1016/j.neuroimage.2006.01.021.

Dominic, J., Bhaskhar, N., Desai, A.D., Schmidt, A., Rubin, E., Gunel, B., Gold, G.E., Hargreaves, B.A., Lenchik, L., Boutin, R., Chaudhari, A.S., 2023. Improving data-efficiency and robustness of medical imaging segmentation using inpainting-based self-supervised learning. Bioengineering 10. doi:10.3390/bioengineering10020207.

Ferrari, A.J., Stockings, E., Khoo, J.P., Erskine, H.E., Degenhardt, L., Vos, T., Whiteford, H.A., 2016. The prevalence and burden of bipolar disorder: findings from the global burden of disease study 2013. Bipolar Disorders 18, 440–450. doi:10.1111/bdi.12423.

Fischl, B., Sereno, M.I., Dale, A., 1999. Cortical surface-based analysis: Ii: Inflation, flattening, and a surface-based coordinate system. NeuroImage 9, 195 – 207.

Gao, F., Yoon, H., Wu, T., Chu, X., 2020. A feature transfer enabled multi-task deep learning model on medical imaging. Expert Systems with Applications 143, 112957. doi:https://doi.org/10.1016/j.eswa.2019.112957.

Hesamian, M.H., Jia, W., He, X., Kennedy, P., 2019. Deep learning techniques for medical image segmentation: Achievements and challenges. Journal of Digital Imaging 32, 582–596. doi:10.1007/s10278-019-00227-x.

Huang, S.C., Pareek, A., Jensen, M., Lungren, M.P., Yeung, S., Chaudhari, A.S., 2023. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. npj Digital Medicine 6. doi:10.1038/s41746-023-00811-0.

Huntgeburth, S.C., Petrides, M., 2012. Morphological patterns of the collateral sulcus in the human brain. European Journal of Neuroscience 35, 1295–1311. doi:https://doi.org/10.1111/j.1460-9568.2012.08031.x.

Iglesias, J.E., Billot, B., Balbastre, Y., Tabari, A., Conklin, J., González, R.G., Alexander, D.C., Golland, P., Edlow, B.L., Fischl, B., et al., 2021. Joint super-resolution and synthesis of 1 mm isotropic mp-rage volumes from clinical mri exams with scans of different orientation, resolution and contrast. Neuroimage 237, 118206.

Jaiswal, A., Babu, A.R., Zadeh, M.Z., Banerjee, D., Makedon, F., 2020. A survey on contrastive self-supervised learning. Technologies 9, 2.

Jensen, B., 2016. Influence of Maturation, Pathology and Functional Lateralization on 3D Sulcal Morphology using MRI. Ph.D. thesis. Technical University of Denmark, DTU Compute.

Kao, C.Y., Hofer, M., Sapiro, G., Stern, J., Rehm, K., Rottenberg, D.A., 2007. A geometric method for automatic extraction of sulcal fundi. IEEE transactions on medical imaging 26, 530–540.

Klein, D., Rotarska-Jagiela, A., Genc, E., Sritharan, S., Mohr, H., Roux, F., Han, C.E., Kaiser, M., Singer, W., Uhlhaas, P.J., 2014. Adolescent brain maturation and cortical folding: Evidence for reductions in gyrification. PLoS ONE 9, e84914. doi:10.1371/journal.pone.0084914.

Kochunov, P., Mangin, J.F., Coyle, T., Lancaster, J., Thompson, P., Rivière, D., Cointepas, Y., Régis, J., Schlosser, A., Royall, D.R., Zilles, K., Mazziotta, J., Toga, A., Fox, P.T., 2005. Age-related morphology trends of cortical sulci. Human Brain Mapping 26, 210–220. URL: https://doi.org/10.1002/hbm.20198, doi:10.1002/hbm.20198.

Kochunov, P., Rogers, W., Mangin, J.F., Lancaster, J., 2011. A library of cortical morphology analysis tools to study development, aging and genetics of cerebral cortex. Neuroinformatics 10, 81–96. doi:10.1007/s12021-011-9127-9.

Leroy, F., Cai, Q., Bogart, S.L., Dubois, J., Coulon, O., Monzalvo, K., Fischer, C., Glasel, H., der Haegen, L.V., Bénézit, A., Lin, C.P., Kennedy, D.N., Ihara, A.S., Hertz-Pannier, L., Moutard, M.L., Poupon, C., Brysbaert, M., Roberts, N., Hopkins, W.D., Mangin, J.F., Dehaene-Lambertz, G., 2015. New human-specific brain landmark: The depth asymmetry of superior temporal sulcus. Proceedings of the National Academy of Sciences 112, 1208–1213. doi:10.1073/pnas.1412389112.

Lin, B., Ye, F., Zhang, Y., 2021a. A closer look at loss weighting in multi-task learning. arXiv preprint arXiv:2111.10603 .

Lin, H.Y., Huang, C.C., Chou, K.H., Yang, A.C., Lo, C.Y.Z., Tsai, S.J., Lin, C.P., 2021b. Differential patterns of gyral and sulcal morphological changes during normal aging process. Frontiers in Aging Neuroscience 13. doi:10.3389/fnagi.2021.625931.

Liu, X., Song, L., Liu, S., Zhang, Y., 2021. A review of deep-learning-based medical image segmentation methods. Sustainability 13, 1224. doi:10.3390/su13031224.

Lyu, I., Bao, S., Hao, L., Yao, J., Miller, J.A., Voorhies, W., Taylor, W.D., Bunge, S.A., Weiner, K.S., Landman, B.A., 2021. Labeling lateral prefrontal sulci using spherical data augmentation and context-aware training. NeuroImage 229, 117758. doi:https://doi.org/10.1016/j.neuroimage.2021.117758.

van der Maaten, L., Hinton, G., 2008. Visualizing data using t-sne. Journal of Machine Learning Research 9, 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

Mangin, J.F., Frouin, V., Bloch, I., Rogis, J., Lopez-Krahe, J., 1995. From 3d magnetic resonance images to structural representations of the cortex topography using topology preserving deformations. Journal of Mathematical Imaging and Vision 5, 297–318. doi:10.1007/bf01250286.

McConnell, S.K., 1995. Constructing the cerebral cortex: neurogenesis and fate determination. Neuron 15, 761–768.

Millier, A., Schmidt, U., Angermeyer, M., Chauhan, D., Murthy, V., Toumi, M., Cadi-Soussi, N., 2014. Humanistic burden in schizophrenia: A literature review. Journal of Psychiatric Research 54, 85–93. doi:10.1016/j.jpsychires.2014.03.021.

Ochiai, T., Grimault, S., Scavarda, D., Roch, G., Hori, T., Rivière, D., Mangin, J.F., Régis, J., 2004. Sulcal pattern and morphology of the superior temporal sulcus. NeuroImage 22, 706–719. doi:https://doi.org/10.1016/j.neuroimage.2004.01.023.

Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 .

Perrot, M., Rivière, D., Mangin, J.F., 2011. Cortical sulci recognition and spatial normalization. Medical Image Analysis 15, 529–550. doi:https://doi.org/10.1016/j.media.2011.02.008. special section on IPMI 2009.

Puonti, O., Iglesias, J.E., Van Leemput, K., 2016. Fast and sequence-adaptive whole-brain segmentation using parametric bayesian modeling. NeuroImage 143, 235–249. doi:https://doi.org/10.1016/j.neuroimage.2016.09.011.

Robinson, N., Bergen, S.E., 2021. Environmental risk factors for schizophrenia and bipolar disorder and their relationship to genetic risk: Current knowledge and future directions. Frontiers in Genetics 12. doi:10.3389/fgene.2021.686666.

Roell, M., Cachia, A., Matejko, A., Houdé, O., Ansari, D., Borst, G., 2021. Sulcation of the intraparietal sulcus is related to symbolic but not non-symbolic number skills. Developmental Cognitive Neuroscience 51, 100998. doi:https://doi.org/10.1016/j.dcn.2021.100998.

Rohlfing, T., Brandt, R., Menzel, R., Maurer Jr, C.R., 2004. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. NeuroImage 21, 1428–1442.

Salehi, S.S.M., Erdogmus, D., Gholipour, A., 2017. Tversky loss function for image segmentation using 3d fully convolutional deep networks, in: Wang, Q., Shi, Y., Suk, H.I., Suzuki, K. (Eds.), Machine Learning in Medical Imaging, Springer International Publishing, Cham. pp. 379–387.

Schindler, G., Dellaert, F., 2004. Atlanta world: An expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments, in: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., IEEE. pp. I–I.

Shi, Y., Tu, Z., Reiss, A.L., Dutton, R.A., Lee, A.D., Galaburda, A.M., Dinov, I., Thompson, P.M., Toga, A.W., 2007. Joint sulci detection using graphical models and boosted priors, in: IPMI, pp. 98–109.

Thorup, A.A.E., Hemager, N., Søndergaard, A., Gregersen, M., Prøsch, Å.K., Krantz, M.F., Brandt, J.M., Carmichael, L., Melau, M., Ellersgaard, D.V., Burton, B.K., Greve, A.N., Uddin, M.J., Ohland, J., Nejad, A.B., Johnsen, L.K., van Themaat, A.H.V.L., Andreassen, A.K., Vedum, L., Knudsen, C.B., Stadsgaard, H., Jepsen, J.R.M., Siebner, H.R., Østergaard, L., Bliksted, V.F., Plessen, K.J., Mors, O., Nordentoft, M., 2018. The danish high risk and resilience study—VIA 11: Study protocol for the first follow-up of the VIA 7 cohort -522 children born to parents with schizophrenia spectrum disorders or bipolar disorder and controls being re-examined for the first time at age 11. Frontiers in Psychiatry 9. doi:10.3389/fpsyt.2018.00661.

Thorup, A.A.E., Jepsen, J.R., Ellersgaard, D.V., Burton, B.K., Christiani, C.J., Hemager, N., Skjærbæk, M., Ranning, A., Spang, K.S., Gantriis, D.L., Greve, A.N., Zahle, K.K., Mors, O., Plessen, K.J., Nordentoft, M., 2015. The danish high risk and resilience study – VIA 7 - a cohort study of 520 7-year-old children born of parents diagnosed with either schizophrenia, bipolar disorder or neither of these two mental disorders. BMC Psychiatry 15. doi:10.1186/s12888-015-0616-5.

Vivodtzev, F., Linsen, L., Bonneau, G.P., Hamann, B., Joy, K., Olshausen, B.A., 2003. Hierachical isosurface segmentation based on discrete curvature. UC Davis: Institute for Data Analysis and Visualization .

White, T., Su, S., Schmidt, M., Kao, C.Y., Sapiro, G., 2010. The development of gyrification in childhood and adolescence. Brain and Cognition 72, 36–45. doi:10.1016/j.bandc.2009.10.009.

Willbrand, E.H., Parker, B.J., Voorhies, W.I., Miller, J.A., Lyu, I., Hallock, T., Aponik-Gremillion, L., Koslov, S.R., Bunge, S.A., Foster, B.L., and, K.S.W., 2022. Uncovering a tripartite landmark in posterior cingulate cortex. Science Advances 8. doi:10.1126/sciadv.abn9516.

Yang, F., Kruggel, F., 2008. Automatic segmentation of human brain sulci. Medical Image Analysis 12, 442–451. doi:https://doi.org/10.1016/j.media.2008.01.003.

Yang, J., Wang, D., Rollins, C., Leming, M., Liò, P., Suckling, J., Murray, G., Garrison, J., Cachia, A., 2019. Volumetric segmentation and characterisation of the paracingulate sulcus on MRI scans. bioRxiv doi:10.1101/859496.

Yaniv, Z., Lowekamp, B.C., Johnson, H.J., Beare, R., 2017. SimpleITK image-analysis notebooks: a collaborative environment for education and reproducible research. Journal of Digital Imaging 31, 290–303. doi:10.1007/s10278-017-0037-8.

Zeng, D., Kheir, J.N., Zeng, P., Shi, Y., 2021. Contrastive learning with temporal correlated medical images: A case study using lung segmentation in chest x-rays (invited paper), in: 2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD), pp. 1–7.

Zhang, Z., Wang, Y., Gao, Y., Li, Z., Zhang, S., Lin, X., Hou, Z., Yu, Q., Wang, X., Liu, S., 2020. Morphological changes in the central sulcus of children with isolated growth hormone deficiency versus idiopathic short stature. Developmental Neurobiology 81, 36–46. doi:10.1002/dneu.22797.

Zhou, Y., Chen, H., Li, Y., Liu, Q., Xu, X., Wang, S., Yap, P.T., Shen, D., 2021. Multi-task learning for segmentation and classification of tumors in 3d automated breast ultrasound images. Medical Image Analysis 70, 101918. doi:https://doi.org/10.1016/j.media.2020.101918.

Medical Imaging and Applications

Master Thesis, June 2023

3D End-to-End Mesh Reconstruction from Pre-Operative CT

Farahdiba Zarin[a], Vinkle Srivastav[b], Nicolas Padoy[c]

[a] *fzarin@unistra.fr*
[b] *srivastav@unistra.fr*
[c] *npadoy@unistra.fr*

Abstract

Advancements in 3D surface representation have been followed by tremendous achievements in computer vision tasks such as detailed scene reconstruction and view synthesis. The potential these advancements pose in medical imaging has been largely unexplored, primarily due to the lack of detailed organ reconstruction emphasizing surface accuracy. The aim of this thesis is to bring the recent advancements in 3D geometric computer vision to medical imaging by enabling detailed and accurate organ surface reconstruction. First, we explore explicit surface representation in the form of 3D surface mesh reconstruction directly from 3D CT volumes. The explicit representation requires geometrical primitives, such as triangles in this case, for discrete representations of the surfaces. Implicit representations, on the other hand, are continuous in nature as they do not rely on geometric primitives but rather on decision boundaries such as level-sets, and hence are not limited by resolution. Given these benefits of implicit representations, we progress with implicit surface generation for further improvement of the reconstruction. Our contribution can be summarized as follows: **i)** End-to-end implicit 3D surface generation based on occupancy values from sampled query points; **ii)** Sampling technique designed to eliminate reliance on ground truth for training and testing; **iii)** Optimization of implicit pipeline with the new sampling method. Our optimization of the implicit pipeline with the new sampling method resulted in a 20-fold increase of the reported chamfer distance, with resulting predicted surface reconstructions of the liver organ having Hausdorff Distance of 3.671 ± 1.995 mm and Average Symmetric Surface Distance of 1.428 ± 0.716 mm.

Keywords: Surface Reconstruction, Implicit Representation, Mesh Reconstruction, 3D Computer Vision, Geometric Deep Learning, Deep Learning

1. Introduction

Computer vision techniques have recently made significant strides in capturing and analyzing 3D surface information across diverse applications. However, their specific application and potential impact in the medical domain necessitate further exploration and investigation.

Prior works in the computer vision domain focused on rendering objects using explicit representations, though implicit ones have gained more favor due to their ability to represent surfaces continuously without the typical restraints of resolution caused by discrete representations (Tewari et al., 2020). Explicit representations, such as point clouds, triangular meshes, or voxel grids, require geometric primitives for the description of geometric objects as presented in Figure 1. Implicit representations map the object from its 3D space to a continuous domain by defining the surface of the object as the level-set of a function. This can be the zero-level set for the signed-distance functions (Park et al., 2019), or occupancy values (Mescheder et al., 2019). Due to triangular meshes being most commonly used by rendering tools, most other representations are converted into meshes. The implicit representations are often rasterized into explicit representations and converted into meshes by marching cubes for visualization.

While explicit representations were common, they are significantly limited by memory requirements, which increase greatly with the spatial resolution, growing cubically. Additionally, these methods do not offer accurate representations of surfaces, requiring higher

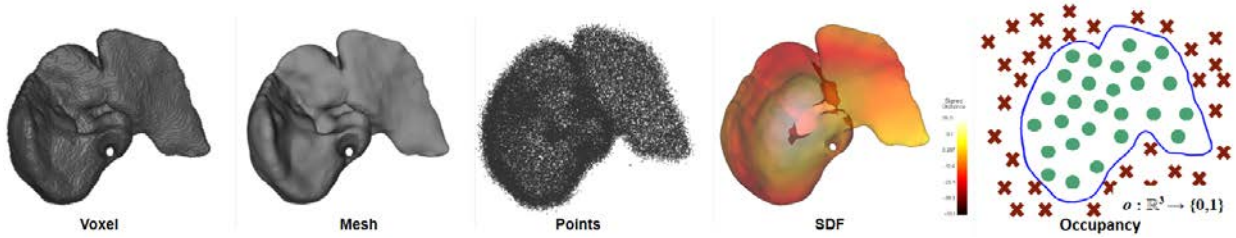


Figure 1: Explicit representations of voxels, mesh, and point cloud. The implicit representation of SDF is presented as the opposite of the functional representation, hence why the distance of the object surface from a bounding box forms the heatmap values with the bounding box as the zero level-set. Typically points inside the object are negative and outside are positive, with the magnitude increasing with the distance from the surface. Occupancy is an implicit representation that bases the surface as the decision boundary, where points inside are assigned 1 and points outside are 0.

resolution than is practically possible in order to represent surfaces in their necessary details. This formed the major limitation in the representation of large objects of scenes, as sufficient detail was difficult to capture.

The use of implicit representation overcomes the limitations of explicit representations and has thus revolutionized the field by producing works such as neural rendering for photorealistic view synthesis of scenes (Mildenhall et al., 2021), generalized shape reconstruction of objects and their inner structures (Chibane et al., 2020b), high-resolution digitization of clothes humans (Saito et al., 2019). All of these works, and more, have the characteristic of detailed and accurate surface rendering.

In this work, we aim to leverage these advancements in computer vision within the medical field, particularly for enhancing the understanding and diagnosis of diseases through comprehensive organ surface analysis. We focus on the explicit and implicit techniques for liver mesh reconstruction using pre-operative CT and MRI images, with a particular emphasis on preserving the intricate details of the reconstructed organs' surfaces.

We investigate and compare two approaches for 3D mesh reconstruction directly from CT and MRI images in an end-to-end manner. Since most rendering tools process meshes, our first approach is based on Voxel2Mesh (Wickramasinghe et al., 2020), which utilizes a graph-based mesh decoder network to directly reconstruct the mesh representation from a 3D CT image. The second approach involves an implicit pipeline utilizing IF-Nets (Chibane et al., 2020a), which decodes the implicit surface representation using features learned from the input image.

The explicit pipeline in its current form suffers from a notable limitation, as it tends to produce overly smoothed surfaces as outputs, thereby compromising the preservation of fine surface details. This smoothing effect can lead to the loss of critical information necessary for accurate organ reconstruction, and attempting to recover finer details can produce noisy meshes. On the other hand, the implicit pipeline, while capable of restoring surface details for already reconstructed objects, does not fully leverage the potential of utilizing learned features from 3D images directly for recovering such intricate details and instead heavily relies on prior knowledge of the target object by taking pre-computed binary segmentation masks as inputs.

In our research, we aim to address these limitations by proposing ideas for both the explicit and implicit pipelines. For the explicit pipeline, we explore techniques to mitigate the issue of over-smoothing and improve the preservation of surface details during the reconstruction process while suppressing the noisy artifacts introduced in the mesh generation process.

In the case of the implicit pipeline, we exploit its capabilities of reproducing highly detailed surfaces from learned features by directly incorporating the information from the 3D images into the detail recovery process. We do so with careful optimization of the learning, as the network is prone to over-fitting in plenty of cases by virtue of its capabilities in learning detailed surfaces. We also utilize rigorous pre-processing methods to ensure spatial alignments, which serve as an important step in fully utilizing the abilities of the implicit feature decoder. This enables us to produce more comprehensive surface reconstruction bypassing the intermediary step of obtaining a coarse object reconstruction for the IF-Net to work, making the process an end-to-end one completely trainable from scratch.

These adaptations of the explicit and implicit pipeline for retaining more surface information during organ reconstruction offer immense potential for conducting detailed surface analysis of reconstructed organs for disease diagnosis and progression assessment. One such downstream application is the liver surface nodularity (LSN) computation to diagnose liver cirrhosis. The diagnosis, in this case, is primarily invasive, requiring biopsy for confirmation as the gold standard (Ginès et al., 2021). Non-invasive alternatives utilizing pre-operative CT or MRI images have emerged in recent years, possessing the potential to replace the invasive procedures. LSN is one such measure which provides an indication of the degree of cirrhosis (Catania et al., 2021; Elkassem et al., 2022; Smith et al., 2016).

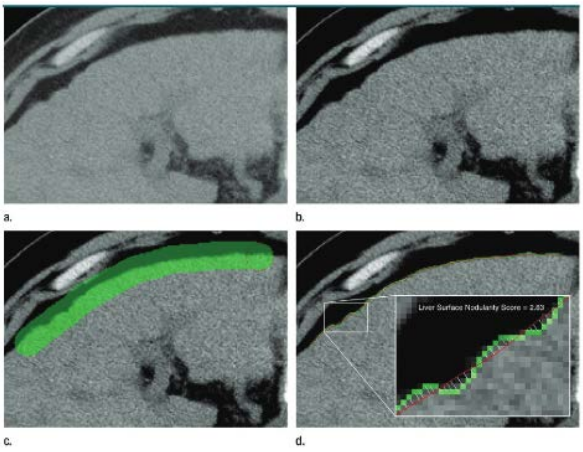


Figure 2: Semi-automated approach of computing liver surface nodularity (Smith et al., 2016).

By analyzing the irregularities or abnormalities on the outer surface of the liver, LSN can provide valuable information for assessing the extent of cirrhosis without the need for invasive interventions. Currently, the assessment of LSN relies on manual or semi-automated methods. These approaches, including those that employ statistical evaluation or machine learning techniques Kim et al. (2020); Kotowski et al. (2023), commonly utilize a limited number of 2D slices for analysis. The typical methodology begins with outlining the external surface of the liver on a 2D slice, which serves as a reference for the location from which the LSN is subsequently measured (Figure 2. Despite relying on the boundary contour for quantifying LSN, these methods often neglect to fully leverage the 3D surface information for more comprehensive measurements. This utilization of 3D surface information is constrained by the scarcity of automated tools capable of directly reconstructing organs from CT or MRI images while prioritizing the preservation of surface details.

Our main contribution therefore can be summarized as follows:

- End-to-end implicit pipeline to provide high-resolution organ reconstruction directly from the 3D images.
- Sampling strategy independent of the ground truth for enforcing non-reliance of ground truth at test time.
- We further showcase the ability of the pipeline to generalize on images from different sources with no predefined knowledge of the location of the target organ by making our pipeline applicable on the diverse TotalSegmentator (Wasserthal et al., 2022) dataset.
- Improvement of end-to-end explicit pipeline to obtain detailed organ surface reconstruction.

By further elaborating on the implicit pipeline and performing a comparative analysis with the explicit pipeline, we aim to overcome the limitations of previous methods and achieve more accurate and detailed organ surface reconstructions.

2. State of the art

Most surface reconstructions in computer vision revolve around representations that can be categorized as either explicit or implicit. Explicit surface representations, including voxel grids, point clouds, and meshes, involve the discretization of the surface into distinct elements with each representation having their own advantages and disadvantages. Implicit surface representations such as SDFs and occupancy functions have the capacity to represent surfaces continuously, and have thus gained much traction in the recent years. The subsequent sections provide a detailed explanation of these various representations and their architectures in organ reconstruction. While extensive work exists in 3D reconstruction from single-view and multi-view 2D images (Fu et al., 2021; Han et al., 2019), we emphasize more on 3D surface reconstruction from 3D images, in particular use case in organ reconstruction.

2.1. Voxels

Voxelized representations are the most common discrete representations found in 3D medical segmentation. In this case, the entire object is the form of 3D voxels occupying a 3D box grid in locations where there the object exists. Surface representations using voxelized outputs are not typical in medical images, as other forms receive more preference. Typically, the resolution is a huge constrain in these sort of representations, as accurate representation of the object and its surface boundary is possible with higher resolution. However, increasing the resolution dramatically increased the storage allocated, as well as processing of voxelized outputs are computationally expensive.

2.2. Points

Point clouds are a representation of an object using numerous unconnected points in the 3D coordinate space. However, sparse point clouds often lack detailed representation, and increasing the number of points to capture more details significantly increases the required storage space for data. The absence of connectivity between points significantly limits the information contained by point clouds, and does not allow closed surface representations. Most medical imaging tasks using point clouds focus on segmentation, with emphasis on surface points. PC-UNet (Ye et al., 2021) utilizes PointNet (Qi et al., 2017) for point cloud reconstruction of

cardiac walls and for further refinement of their segmentation output. Balsiger et al. (2019); Cai et al. (2019); Yao et al. (2020) integrate point cloud reconstruction into their pipeline for the same purpose.

2.3. Meshes

A mesh is a representation that discretizes surfaces or volumes in three-dimensional space (Hoppe et al., 1993). It comprises interconnected geometric elements called vertices, edges, and faces, which collaboratively determine the object's shape, topology, and characteristics. Specifically, in a triangular mesh, these elements form triangles, with three vertices and three edges forming each triangle. The vertex positions are defined by the 3D coordinates of the points in space, and each vertex is connected to one other points by their edge. 3 such neighbouring vertices connected by their edges to form an enclosed 2D plane, which is the triangular face.

Whole organ mesh generation from 3D CT or MRI image in Voxel2Mesh (Wickramasinghe et al., 2020), MeshDeformNet (Kong et al., 2021), and Vox2Cortex (Bongratz et al., 2022) follow similar methodology. The architectures contain 3D CNN Encoder-Decoder that are UNet variants for feature extraction, and Graph Convolution Networks as the mesh decoder, based on the initially proposed Pixel2Mesh (Wang et al., 2018) for mesh generation from 2D RGB images. The mesh decoder utilizes features obtained from the CNN backbone to deform a initial, usually spherical template mesh, through a series of deformation blocks. This is made possible by the treatment of the meshes as graphs where the vertices are the nodes and the faces are the connections. Each blocks receives features from a different resolution through spatially correlated unpooling of features, allowing both global and local learned features to be used. The features are extracted from both the encoder and the decoder of the CNNs by MeshDeformNet and Vox2Cortex, while Voxel2Mesh only extracts decoder features.

Voxel2Mesh reports more generalization, while the latter two emphasize and evaluate on one specific organ for their design. One key component in the pipeline are the use of initial meshes which are aligned in either shape (Vox2Cortex using smoothed version of the ground truth organ meshes), or spatially (MeshDeformNet spatially align their spheres first to correspond to the organ location), which limits the scope of generalization on different datasets.

Additionally, the loss functions used, while all have mostly similar components, treat the mesh regularization factors in the loss differently to suit their respective purposes. Voxel2Mesh and MeshDeformNet increase the coefficients of the regularization terms to promote smoother surfaces, which generalizes on the generated meshes to the extent of eroding all surface details. Vox2Cortex use different empirically found the coefficients for each organ, reporting different optimums for each organ. The key component reported by them is the curvature-weighted chamfer loss, which allows more accurate surface reconstruction by emphasizing on the curvature of the surface. One major disadvantage of meshes is that while they provide information regarding connectivity of the surface points, they can self-intersect, or the architectures may only allow working with limited resolution (Liao et al., 2018). Typically generation of mesh from other forms of 3D representations use marching cubes (Lorensen and Cline, 1987), but this can induce stair-case artifacts which require postprocessing for removal, hence making direct mesh generation a more desirable approach.

2.4. Signed Distance Functions

Signed distance functions implicitly represent the location of a point on the surface as the zero level set, and represent all other points with respect to the zero level set surface (Park et al., 2019). In DeepSDF, the function assigns a signed distance value to each point in space, indicating its proximity to the surface by the decreasing magnitude, and represents the location of the points as either inside or outside the surface by assigning negative or positive sign, and outputs zero to represent points exactly on the surface. The architecture takes a 3D coordinate (x,y,z) as input and outputs the corresponding signed distance value. At inference the shape's surface is estimated by querying the network for signed distance values at different points in space. The continuous nature of the SDF representation allows for high-quality surface reconstruction, even in regions with complex geometry or limited data, though DeepLS (Chabra et al., 2020) and SIRENs (Sitzmann et al., 2020) for mesh reconstruction have further improved upon the level of details.

Works of Dangi et al. (2019); Navarro et al. (2019); Xue et al. (2020) apply SDF generation to monitor 3D organ segmentation outputs, with reported improvements compared to other widely used methods that do not take shape into consideration. Liu et al. (2022) notes the necessity of regularization for the SDF loss function to work, and monitor SDF alongside region and pixel-based loss functions for regularization. Xue et al. (2020) reports unstable learning when solely using the L1 loss from DeepSDF (Park et al., 2019) for the SDF regression, and thus makes further modifications for the loss for stabilization. The architectures used in these works are all primarily UNet-based CNNs, and differ drastically from that proposed in Park et al. (2019) for SDF formulation. In the DeepSDF framework, at the start of training, each data point is assigned a randomized token as the latent vector. These latent vectors and coordinates of the data points are inputs of the autodecoders. At training time, these latent vectors along with the decoder weights are updated simultaneously. At inference phase, the optimal latent vector is estimated for each data point by the decoder.

2.5. Occupancy Functions

Occupancy functions use deep features of query points of a 3D object to identify if the point belongs inside or outside the object, and is essentially a form of binary classification first introduced for 3D reconstruction from 2D images, Point Clouds, and 3D images in Occupancy Networks (Mescheder et al., 2019). This is achieved in the continuous domain by feature sampling from query point features to make predictions of the occupancy for not only the existing surface points, but also other possible surface points around the surface, primarily obtained by dense sampling of query points from the object, either uniformly throughout the cubical bounding box containing the object (Mescheder et al., 2019) or concentrated around the surface boundary (Chibane et al., 2020a). These sampling strategies also makes 3D shape reconstruction at resolutions higher than the input image possible. For 3D images, the typically architecture follows a 3D CNN encoder for feature extraction from the image, and fully-connected layers as the decoder for classification of the points.

In order to eliminate any level of dependency on the location of the query points themselves and emphasize learning from features only, IF-Nets utilized only the features of the query points themselves to obtain their occupancy, and excessively sample around the query points for obtaining the features shared by those neighbouring points to promote learning in the continuous domain (Chibane et al., 2020a).

Khan and Fang (2022); Marimont and Tarroni (2022) both extend the use of 3D CNN encoders with Fully-Connected implicit decoders for implicit 3D medical image segmentation of the lung and pancreas (Marimont and Tarroni, 2022) and head and neck (Khan and Fang, 2022) respectively.

Prior method proposed by Khan and Fang (2022) relies on creating bounding boxes around the target organs for generation of query points sampled densely around the organs. On diverse datasets such as TotalSegmentator (Wasserthal et al., 2022) where all images are not centered around the same organ, this requires careful crafting of bounding boxes and prior knowledge of the organ location. Marimont and Tarroni (2022) also relies on the ground truth by generating sampled points from the ground truth segmentation mask. In order to eliminate influence of the ground truth or prior knowledge in preparation of the input, we propose a sampling strategy to be reliant solely on the input image itself for query point selection to make training data independent of the ground truth. This also ensures non-reliance on ground truth information at test time. Our method ignores the background completely and selectively samples in and around regions containing useful information while overcoming the necessity of the location of the organs being predefined as in (Khan and Fang, 2022; Marimont and Tarroni, 2022). This simple change makes our method applicable on completely unseen images at inference mode, with no existence of ground truth or prior knowledge of the images being required for the reconstruction. It is also non-specific and applicable to all organs visible in the image. We further elaborate on the different methods of processing the learned features to find the most optimal one.

3. Material and methods

3.1. Dataset

We use the TotalSegmentator (Wasserthal et al., 2022) dataset due to its versatility as a large-scale medical image dataset containing segmentation masks for liver as well as other abdominal organs. At the time of retrieval, 1203 usable images existed where 923 had liver ground truth segmentation masks with content. The 3D end-to-end pipelines require inputs to be cubical, which is achieved by padding to match the dimension of the largest image following resampling to uniform voxel spacing. Given the voxel spacing being the same for all images of the TotalSegmentator dataset, the 53 liver images having more than 500 slices along any axis are not included in the dataset to avoid excessive padding of all the other images in the preprocessing. Since images are not guaranteed to be centered on the abdomen, images containing no liver or only fragments of the liver are discarded by thresholding based on the liver quantity present. This threshold is determined by the minimum possible liver volume of 1150 mL, derived from the mean and standard deviation of 1533 mL ± 375 reported by Perez et al. (2022) and the average of 1410 mL ± 271.28 for the CHAOS dataset (Kavur et al., 2020). The resulting 581 CT volumes are split into 291 for training, 116 for validation, and 174 for testing.

3.1.1. Pre-processing

Data preprocessing steps for ground truth mesh and images are generally the same for both the explicit and the implicit pipeline, with some minor differences. Original 500^3 resolution meshes are generated from the padded 3D smoothed versions of the voxelized ground truth segmentation volumes. Each of the padded images undergo intensity normalization with the image mean and standard deviation, following which the images and voxelized ground truth segmentation volumes are resampled to the required resolution of 128^3. The vertices of the ground truth mesh are normalized using the original resolution for the implicit network, which is followed by scaling of the mesh to the target resolution of 256^3. The explicit pipeline differ in that the meshes used are generated from the surface points of the corresponding 128^3 voxelized ground truths after undergoing the same data augmentation process that is applied to the images during training. Finally, vertex normalization is performed.

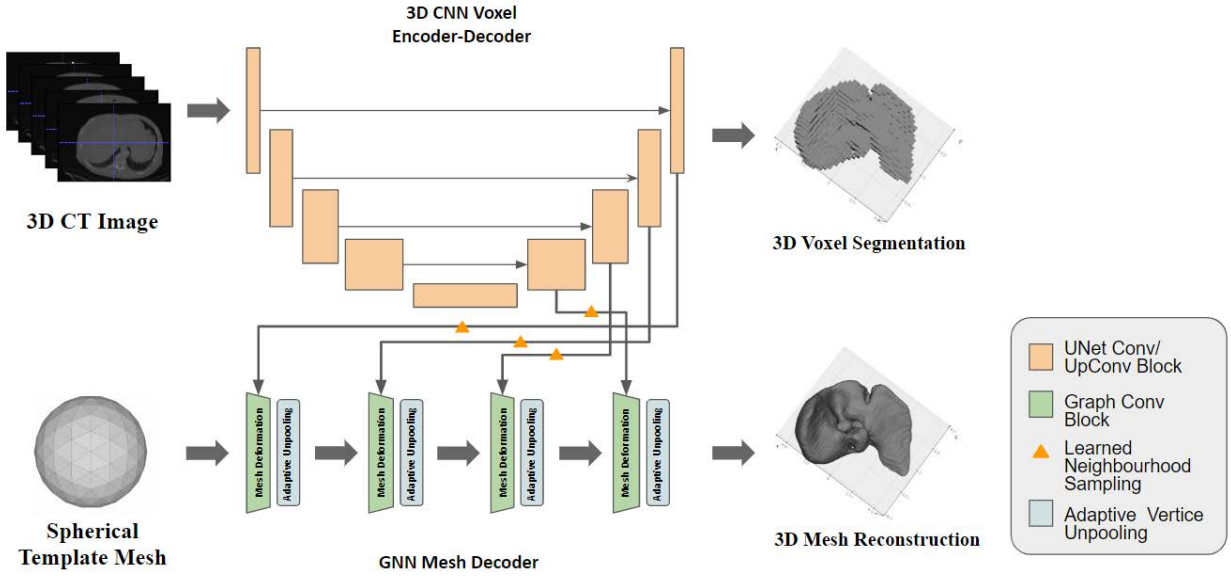


Figure 3: Explicit pipeline for mesh deformation using Voxel2Mesh where the 3D CNN backbone is either a UNet or a R2UNet, and feature extraction is performed from the encoder or decoder or from both and merged. Learned neighbourhood sampling is done to extract features from the corresponding spatial locations by considering a learned neighbourhood around the corresponding vertices. Adaptive unpooling selectively unpools and retains vertices that have been sufficiently deformed by the mesh deformation module prior to it.

3.2. Architecture

3.2.1. Explicit

The explicit pipeline used is Voxel2Mesh (Wickramasinghe et al., 2020) for deforming a spherical template mesh using a graph neural network based mesh decoder. As seen in Figure 3, features for basing the deformation are extracted through learned neighbourhood sampling from a voxel encoder-decoder CNN backbone. The CNN backbone used in Voxel2Mesh is a traditional 3D CNN based UNet (Çiçek et al., 2016). In Voxel2Mesh, features from the decoder part of the UNet are spatially sampled for the respective vertices. The 3D voxel encoder and decoder and the mesh decoder all consist of 4 blocks. Vertices are added prior to their input into the mesh decoder. Unpooling of vertices occurs following mesh deformation, and the adaptive mesh unpooling selectively unpools only the vertices that underwent sufficient deformation. This degree of deformation is based on the distance of the newly deformed vertices from their parent edges.

The 3D CNN backbone in Figure 3 is trained jointly in parallel with the mesh decoder. Training of the voxel encoder-decoder is monitored using standard cross-entropy loss based on the segmentation mask predicted by the UNet backbone by comparing it with the ground truth segmentation mask generated downsampled to be of the same resolution.

$$L_{Total} = L_{CE} + \lambda_1 \cdot L_{CD} + \lambda_2 \cdot L_{norm} + \lambda_3 \cdot L_{lap} + \lambda_4 \cdot L_{edge} \quad (1)$$

Training of the mesh decoder takes the quality of the mesh into consideration as well to avoid mesh intersection and excessive deformation at early stages. In the equation 3.2.1, L_{CE} is the segmentation loss, and L_{CD} is the chamfer loss computed between all the predicted vertices of a mesh compared with the surface points sampled from the ground truth mesh.

The quality of the output mesh is regulated by the regularization terms, L_{norm}, L_{lap}, and L_{edge} respectively, which all contribute to smoothing and non-intersection of the mesh, as well as suppress artifacts caused by excessive deformation of vertices during mesh generation. L_{norm} is the normal consistency loss that maintains consistency between normals of neighbouring faces, and ensured smoothness since the loss is lowest when the faces are parallel. L_{lap} promotes fewer self-intersections by applying another variant of smoothing based on the normals, and L_{edge} specifically works to ensure all edges are of uniform length, and hence do not deform much from each other. The total of all of these losses contribute to the training of the entire architecture.

3.2.2. Implicit

If-Net based architectures are employed to construct the occupancy function, representing the implicit form of the reconstructed surface. These architectures are designed to learn and infer the occupancy status of points in three-dimensional space, enabling the implicit representation of the surface.

The baseline architecture adopts a 3D UNet to generate a voxelized segmentation of the organ, as in order for the IF-Nets to work, a segmentation mask is needed.

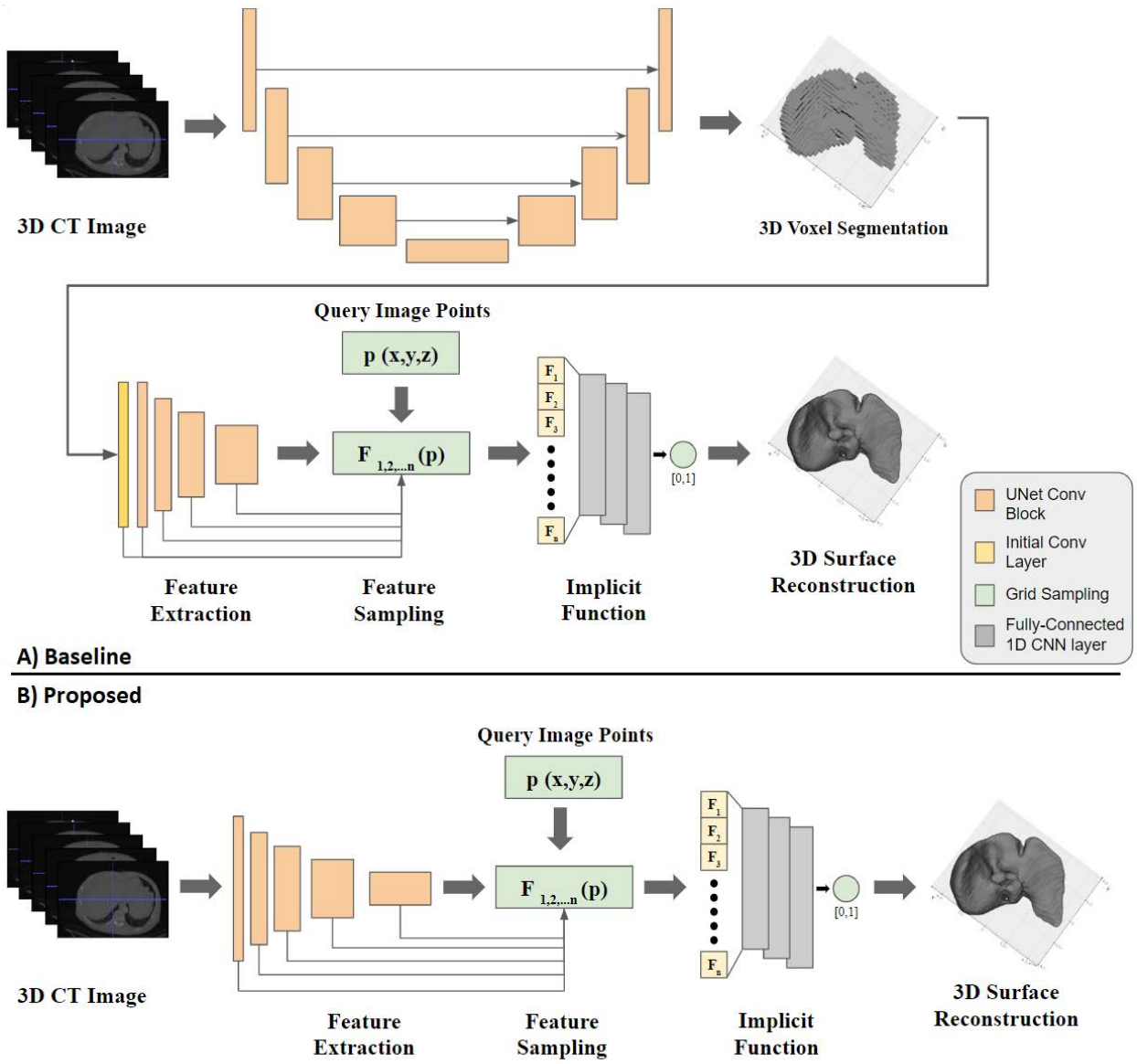


Figure 4: **A)** Baseline pipeline for generating occupancy function from voxelized 3D UNet outputs. **B)** Implicit pipeline for occupancy generation from CT volumes in an end-to-end manner. The feature extraction from the image is performed by a 3D CNN UNet encoder for **B)**. These features are sampled for the input query points and further interpolated for additional arbitrary points surrounding the query points by the feature sampling stage in both pipelines, and the features for each corresponding point processed by fully connected layers in the IF-Net decoder to output an occupancy value for all the query as well as the additional arbitrary points.

This voxelized object serves as the input for IF-Nets to generate the occupancy function. The mesh generation process, performed in the inference mode, employs marching cubes to convert the occupancy values into a surface mesh representation.

During the training process of the 3D UNet, the goal is to minimize the standard cross-entropy loss. This loss is computed based on the predicted segmentation mask value for each voxel. The network is trained to minimize this loss, aiming to improve the alignment between the predicted segmentation and the ground truth. Outputs of the 3D UNet are then used as input for the IF-Nets to generate the occupancies. The pipeline hence relies on training a 3D UNet network for the voxel segmentation first, followed by training of the IF-Net afterwards.

The IF-Nets employed in the implicit pipeline are designed to process inputs of dimension 128^3 and generate higher resolution meshes of 256^3. This particular network architecture was introduced in the study conducted by Chibane et al. (2020a). In the baseline approach from Figure 4 **A)**, the first method of sampling query points from the ground truth mesh is utilized, and all parameters follow the default settings presented in the original

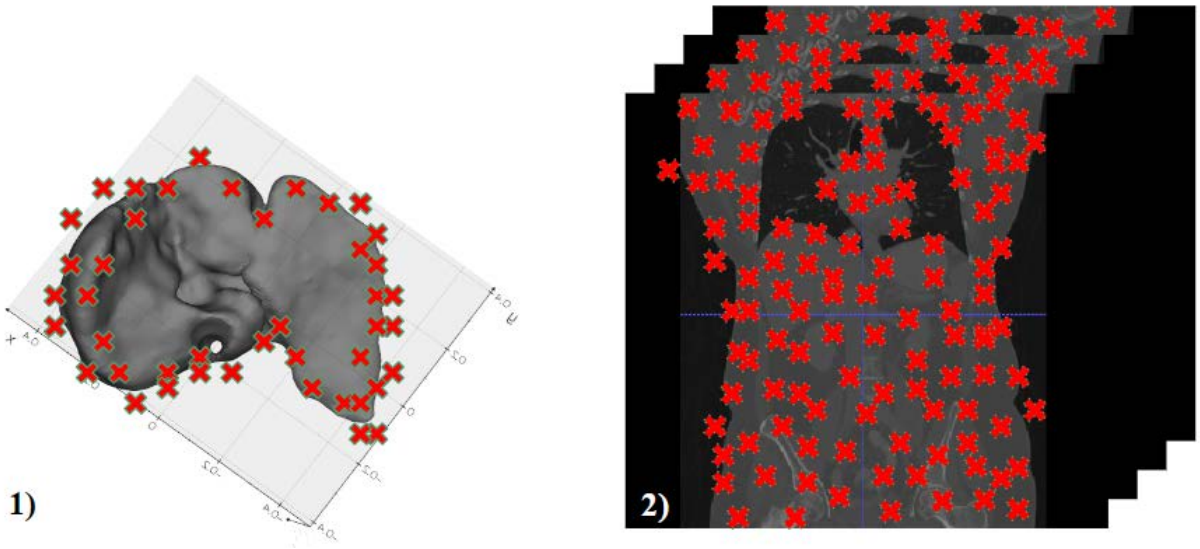


Figure 5: Sampling method for query point generation of for the implicit pipeline. 1) Points are generated from the ground truth mesh, necessitating the presence of a ground truth and the influence of it. 2) Sampling is performed directly from the image itself while still discarding excessive background points with an intensity threshholding designed to sample in and around the organs and tissues.

work for voxel super-resolution using inputs of resolution 128^3. Training of the IF-Nets uses standard binary cross entropy loss to measure the disparity between the predicted occupancies and the ground truth occupancies for the input query points. The network takes a 3D voxel volume as input alongside the list of query points for which to generate occupancy values for. The features used for assigning occupancy values to the query points is computed by the learned feature extraction operations performed by the IF-Net encoder on the IF-Net input. The input itself is also included in the list of features. The encoder used here by the IF-Net is a series of 3D convolutional operations, designed to extract features at both the global as well as the local scale. Features are gridsampled to obtain the relevant ones belonging to not only the query points, but also additional arbitrary points surrounding the query point at a distance of the default 0.35 in different directions. This training is not performed end-to-end, as the UNet is trained first to generated the 3D inputs for the IF-Nets.

The second architecture in Figure 4 **B)** eliminates the step of generating voxelized segmentation masks from the images. Instead, it directly takes the 128^3 3D images as inputs along with the query points list, and produces occupancy values to minimize the binary cross entropy loss between the predicted and actual occupancy values. A standard 3D CNN UNet encoder extracts global and local learned features from the images, and the features spatially sampled from the grid for the query points as well as the generated additional points are processed by the IF-Net decoder. The IF-Nets for the selected input resolution contain 3 fully connected convolution layers and an output layer for prediction of the probability of the query point belonging inside the organ as the decoder. This proposed architecture is trained end-to-end.

$$f(\mathbf{p}) : F_1(I_\mathbf{p}) \times F_2(I_\mathbf{p}) \times \ldots \times F_n(I_\mathbf{p}) \rightarrow [0,1] \quad (2)$$

The occupancy [0,1] defined at 3.2.2 is predicted by the deep learning network $f(\mathbf{p})$ for a given point $\mathbf{p} \in \mathbb{R}^3$ is a function of the features $F_1(I_\mathbf{p}) \times F_2(I_\mathbf{p}) \times \ldots \times F_n(I_\mathbf{p})$ extraction from the image I at query point locations $\mathbf{p}$ in the image.

$$L_{BCE}(\mathbf{y}, y) = -\frac{1}{b \cdot n} \sum_{i=1}^{b} \sum_{j=1}^{n} (\mathbf{y}(i) \cdot \log(\sigma(y(i,j))) + (1 - \mathbf{y}(i,j)) \cdot \log(1 - \sigma(y(i,j)))) \quad (3)$$

The sigma (σ) function is applied element-wise to the predicted probabilities of the occupancy $y(i,j)$ being 1. The outer summation iterates over the batch size (b), while the inner summation iterates over the number of total query points, meaning the sampled ones as well as the arbitrary generated ones (n). The binary cross-entropy loss is computed for each element in the batch and query points, and then averaged over the total number of elements ($b \cdot n$).

3.2.3. Sampling

Explicit.

Vertices of the ground truth meshes are sampled according to the number specified for training.

Implicit.

Method 1: Following the IF-Nets data preprocessing, 10k points are sampled from surface of the scaled ground truth mesh. These points are subjected to displacement using distances generated from a normal distribution. The default sigma values used are 0.1 and 0.01, ensuring that the final points are positioned either inside or outside the object and not all at the boundaries. Occupancy values for these points are subsequently generated as per the original methodology described in IF-Nets (Chibane et al., 2020a).

Method 2: To eliminate any reliance on the ground truth for generating input points for the IF-Nets, the specified number of points are sampled non-uniformly from the preprocessed images. First, only points that do not have the background intensity value are selected. The subsequent processing steps remain consistent with the first method. In cases where the number of points meeting the selection criteria is lower than the desired number of points, all eligible points are included.

The number of iterations determines the number of times displaced points generated from the initially selected points. The final number of query points are hence the sampled points times the number of iterations.

3.3. Experimental Setup

3.3.1. Explicit

The voxel2mesh is used as the basis for the explicit pipeline, where default parameters available are used as a baseline for 10k sampled points. Different factors of the regularization term L_{norm} are tested with different feature sources. In the experimental setups, features from the only the encoder, only the decoder, and both the encoder-decoder are passed to the mesh decoder phase to test the best source of 3D CNN features. Additionally, the 3D CNN backbone is replaced with a 3D R2UNet Alom et al. (2018) backbone for further testing the effects of having a reportedly stronger 3D CNN with recurrent residual convolutional blocks as a feature extractor. The number of points sampled from the ground truth for computing the chamfer loss L_{CD} are also varied, with 3k points and 100k points being tested. The default values of λ_1=1, and $\lambda_2 = \lambda_3 = \lambda_4 = 0.1$ are used in the loss function in equation 3.2.1. The impact of regularization is reduced to decrease the degree of smoothing induced by the original pipeline.

Evaluation on the validation set is done for every 1000 iterations, and the best model is selected based on the the lowest chamfer distance is, with the Jaccard Score also being computed for additional monitoring.

3.3.2. Implicit

The baseline UNet architecture with batch normalization was trained first with Adam optimization and a learning rate of 1×10^{-4}. Based on the lowest validation loss, the optimum model is selected and saved. The weights of this pretrained model are used for the frozen layers of the UNet backbone for the baseline. The voxelized outputs generated by the frozen UNet backbone are used as input for the IFNet architecture designed for superresolution of inputs of resolution 128^3 to generate outputs of occupancy in resolution 256^3.

For the end-to-end pipeline, optimization is done using different learning rates for different modules of the architecture, as well by trying different normalization for the 3D CNN encoder. The final learning rates are 1×10^{-4} for the 3D CNN encoder, and 1×10^{-8} for the IF-Net decoder, with Adam optimizer. ReLU activation was used for all layers, and max pooling with kernel size of 2 for the encoder. All reported results were obtained by training with batch size of 2. Different number of sampled points are tested for the sampling done using method 2, while the default IFNet setting of using 10,000 sampled points is kept for points sampled from the ground truth using method 1. The hidden dimensions in the 3 fully-connected layers in the decoder are tested, with the default of 256 as well a higher value of 512 being tried for the different number of points sampled.

3.4. Evaluation

The best model for each experiment is selected based on the performance of the model on the validation set during training. Evaluation on the test set is carried out afterwards using the saved checkpoints. The Jaccard Score and Chamfer Distances are reported for the resolution that the output meshes are generated. Meanwhile, for computing the 90% Hausdorff Distance and the Average Symmetric Surface Distance, the meshes are first converted to the original resolution by reversing all the preprocessing performed on the ground truth meshes. The vertices are then unnormalized and the scaling reversed based with the voxel spacing of 1.5 mm being taken into context. This is to ensure the computed HD and ASSD are in mm.

4. Results

The quantitative and qualitative results of the different pipelines tested and outlined in the previous sections are illustrated below. Typical time taken for the explicit models to train at the resolution reported is 1-2 weeks (longer for R2UNet) on NVIDIA® V100 GPUs, and 2-3 days for the implicit pipeline on NVIDIA® GeForce RTX 3080. Table 1 reports the results of the explicit pipeline, and Table 2 the evaluated results of the implicit pipeline. For the explicit pipeline, the lowest HD, ASSD, and IOU was for the R2UNet 3D CNN backbone, with a lowered regularization of $\lambda_2 = 0.001$ for minimizing output mesh smoothing. The lowerst CD however, was reported by the original voxel2mesh pipeline. Based on the other scores of HD

8.477±13.998, ASSD 4.031±5.097, and IOU 88.92, the explicit pipeline with R2UNet backbone, lower regularization, and the decoder as the feature source is selected for comparison with the implicit pipeline.

Additional qualitative comparison of the use of only decoder features versus using both encoder and decoder features are displayed in Figure 6. The original pipeline exhibits overall smooth meshes, while the other two pipelines with low regularization capture more surface details similar to the ground truth. The chamfer distance is better for some images in the different pipelines, with visible artifacts remaining for the image in the first row for both.

Figure 7 showcases the outputs of 3 images for the original voxel2mesh pipeline, the pipeline with decoder features and low regularization, and the pipeline where the backbone was replaced with a R2UNet. Whole the average HD, ASSD, and IOU are improved for the R2UNet and artifacts were resolved for the first image, the HD are lower for the other two images.

The end-to-end implicit pipeline performed better based on initial experimentations following the optimization, and further trials to obtain the best number of sampled points and decoder hidden dimensions resulted in HD of 3.671±1.995, ASSD 1.428±0.716, IOU 86.15, and CD of 0.0451×10^{-3}. This was obtained for sampled points of 20k with the image as the source of the sampling, and hidden dimensions of 512 for the decoder.

Visually, the end-to-end pipeline performance is compared to the baseline in Figure 8, where the predicted mesh **b)** follows the UNet output **a)**. In the end-to-end pipeline, the predicted mesh **c)** has a shape more similar to the ground truth **d)**, with none of the staircase effects of the base pipeline, and with the lower length of the liver being captured more.

Qualitative and quantitative results of the results of optimization are illustrated in Figure 9, where all the architectures managed to capture the liver in details, but some contained major artifacts. Reducing the decoder learning rate 100 fold removed most of the artifacts, and the consequent changes of switching from instance normalization to batch normalization for the 3D CNN encoder, and then sampling points from the images, all decreased the artifacts further to improve the reported metrics.

Overall, the HD and ASSD which are reported on the same scale of mm are much lower for the implicit pipelines where sampling is done from the images. In comparison, the CD, HD, and reported ASSD are much higher for the best explicit pipeline. Figure 10 and Figure 11 report the HD of 6 images (3 with comparatively better HD and 3 with worse HD) for the best explicit and implicit pipelines. Not much difference is evident visually for the predicted meshes of both pipelines, but the difference is drastic for more difficult images that have higher HD for both pipelines. Huge artifacts, indicated by arrows, exist in some of the meshes predicted by the explicit pipeline as seen for images Figure 11 **iv)** and 11 **v)**, while some parts are some chunks are missing or not captured as seen for image 11 **vi)**.

5. Discussion

Based on the comparisons outlined in section 4, the implicit pipeline outperformed the explicit pipeline after optimization. The explicit pipeline produced major artifacts in the mesh generation even for the best pipeline using the R2UNet, and this was completely resolved using the implicit network.

From Table 1, using decoder versus encoder+decoder features did not produce a major difference in HD, ASSD, and IOU as observed when comparing the results in row 2 with 3, 5 with 6, and 10 with 11. The most drastic improvement was with the reduction of the λ_2 from 0.1 to 0.001, which was followed by drastic improvements in the HD, ASSD, and IOU as seen in all the rows with 0.001 for λ_2. This was made certain by the R2UNet having an improvement of the IOU increasing to 88.92 compared to the 84.58. The CD metric in comparison is not consistent with the improvements reported by the other metrics, since it is lowest for the baseline despite all the other quantitative metrics being among the worst 3. This makes the metric unreliable in selecting the final pipeline, as the metric appears to reward generic smoothing of meshes more heavily, and is also why the pipeline with the smoothed mesh has some of the lowest standard deviation in the HD and ASSD.

While further improvements might have resulted with testing different factors of the regularization term, this is not optimal as too many regularization terms exist for the explicit pipeline where direct meshes are produced. Comparatively, the learning of the implicit pipeline is far simpler, as the point-based approach mean that only binary cross entropy loss is sufficient to monitor the predicted occupancy.

The major advantage of the architecture and learning being simpler for the implicit pipeline mean easier optimization, as not many hyperparameters need to be fine-tuned for the optimum model to be discovered.

The baseline where the UNet is trained beforehand is not optimized further, as Figure 8 makes it evident that the output occupancy is overly reliant on the effectiveness of the UNet, making the IFNet ineffective in this context. The end-to-end pipeline is far more effective as it is not limited by the UNet output, and also has direct access to the features generated from the images while also being trainable in one single step.

The fully-connected decoder of the implicit pipeline has a tendency to overfit and produce artifacts. This is mitigated by tuning the learning rate, with further changes to make the model more generalizable, improving results further. The normalization to be selected was of importance, as artifacts existed for both layer and

Input	Samples	Features	λ_2	3D CNN	HD	ASSD	IOU	CD
128^3	10k	Encoder	0.1	UNet	13.830±11.048	5.828±3.704	80.94	1.136
128^3	10k	Decoder	0.1	UNet	12.189±10.711	5.355±1.985	83.29	**0.660**
128^3	10k	Encoder+Decoder	0.1	UNet	12.746±16.162	5.441±6.892	83.15	1.657
128^3	10k	Encoder	0.001	UNet	12.014±13.917	4.916±3.910	83.69	0.933
128^3	10k	Decoder	0.001	UNet	9.351±11.645	4.421±4.426	86.55	0.809
128^3	10k	Encoder+Decoder	0.001	UNet	9.429±10.367	4.456±4.183	87.02	0.796
128^3	3k	Encoder+Decoder	0.001	UNet	10.798±14.252	5.008±5.227	86.47	1.128
128^3	100k	Encoder+Decoder	0.001	UNet	9.050±12.395	4.612±5.187	87.45	0.964
128^3	10k	Decoder	0.1	R2UNet	11.393±15.018	4.921±5.405	84.58	1.214
128^3	10k	Decoder	0.001	R2UNet	**8.477±13.998**	**4.031±5.097**	**88.92**	0.972
128^3	10k	Encoder+Decoder	0.001	R2UNet	9.110±17.267	4.137±6.130	88.20	1.285

Table 1: Hausdorff Distance (HD), Average Symmetric Surface Distance (ASSD), Jaccard Score (IOU), and Chamfer Distance (CD) of the explicit pipeline. 90% HD and ASSD reported in mm, and CD in x10^{-3}. CD is computed between the sampled ground truth vertices and all of the predicted vertices.

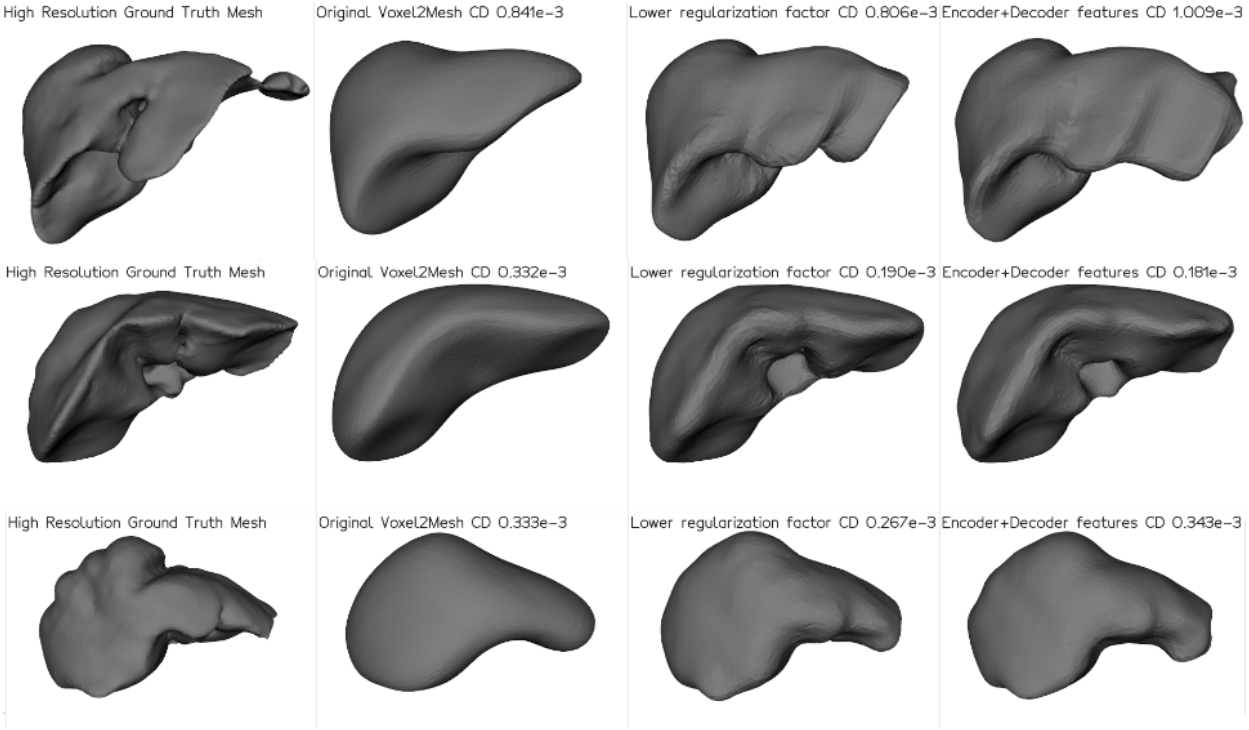


Figure 6: First column from the left indicates high resolution ground truth mesh for several liver examples, with the output of the original voxel2mesh pipeline illustrated in the second column. Results of different feature sources from the 3D UNet backbone (only the decoder versus features from both encoder+decoder) with reduced weightage of normal loss and hence lower impact of regularization factors in the loss function shown in the rightmost two columns. Corresponding chamfer distances computed with respect to the ground truth mesh for the outputs of three pipelines being compared are included in the top.

instance normalization, and these were resolved when we switched to batch normalization. Aside from that leading to the greatest improvement in all the metrics, sampling query points from the image improved the HD from 12.473 to 4.198, and the CD almost 20 fold as it decreased from 1.685×10^{-3} to 0.0731×10^{-3} as seen in Table 2.

We hypothesize this 20-fold improvement in the CD and the remarkable improvements in HD and ASSD due to the alternate sampling method. This could be attributed to the importance of features learned from informative regions, meaning other organs, outside of the target organ in the classification of these outside points to the negative class. We also hypothesize that when the 3D representation of the object already exists, the features extracted by all points inside the object are not that different from each other. However, for a 3D image itself that has the object in a scene surrounded by other objects and a background, the features may appear similar to features extracted from the other objects

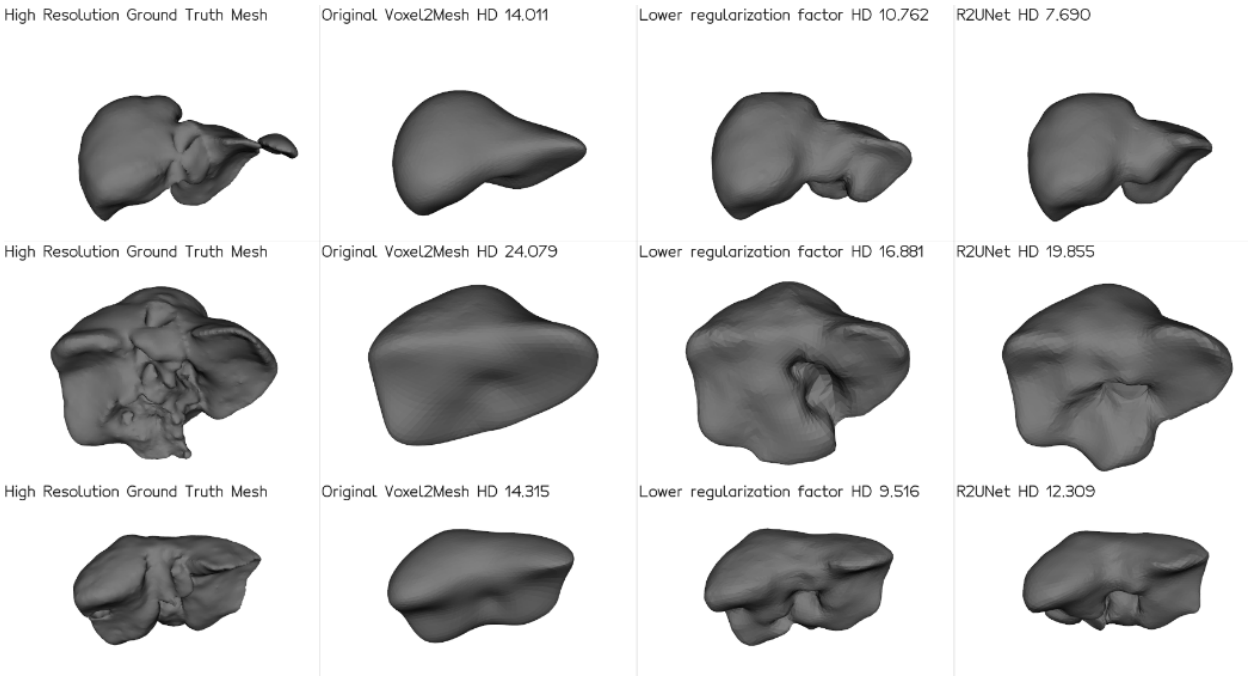


Figure 7: Comparison of hausdorff distance for the original explicit pipeline compared to the pipeline with a lower regularization term in the loss function. Last column to the right indicates the same low regularized pipeline with a R2UNet replacing the backbone 3D CNN voxel architecture.

Network	Sample	Sample Source	Encoder Normalization	Decoder Hidden Dimensions	HD	ASSD	IOU	CD
Baseline	10k	GT	-	256	-	-	75.01	2.120
End-to-End	10k	GT	Layer	256	266.580±15.012	105.036±3.359	02.20	164.86
End-to-End	10k	GT	Instance	256	61.113±79.626	11.527±17.307	82.50	12.609
End-to-End	10k	GT	Batch	256	12.473±35.881	2.873±5.815	85.31	1.685
End-to-End	10k	Image	Batch	256	4.198±6.129	1.557±0.935	85.16	0.0731
End-to-End	20k	Image	Batch	256	3.927±2.887	1.491±0.796	85.35	0.0478
End-to-End	30k	Image	Batch	256	4.266±6.453	1.568±1.531	85.22	0.0831
End-to-End	10k	Image	Batch	512	4.070±5.337	1.512±0.806	85.52	0.0656
End-to-End	20k	Image	Batch	512	**3.671±1.995**	**1.428±0.716**	**86.15**	**0.0451**
End-to-End	30k	Image	Batch	512	3.810±3.024	1.442±0.858	85.78	0.0508

Table 2: Results of the implicit pipeline, with 90% Hausdorff Distance and Average Symmetric Surface Distance reported in mm, and Chamfer Distance in CD x10-3. HD and ASSD not computed for the baseline. Sample source indicates GT for high resolution ground truth mesh, and Image for the input 3D CT volumes.

(other abdominal organs with similar tissue construction in this case), and might be why training is aided by learning of points from surrounding objects and their features. It might also be why careful optimization of training is necessary to ensure learning focuses on correctly identifying the points based on the features and to remove the large amount of artifacts that result.

Further, as the entire image is sampled, increasing the number of points samples from 10k to 20k improved all the reported metrics further. Further increasing the points did not result in improvement, likely due to points repeating too often. While increasing the hidden dimensions of the decoder from 256 to 512 didn't result in any drastic changes in the metrics, the improvement that resulted was still consistent for all metrics, with the lowest standard deviation also being reported for the HD and ASSD with this combination of parameters, indicating a great degree of consistency between the results generated by the network. This makes the implicit pipeline more reliable, as the results of the explicit pipeline were not consistent for all images, since despite improvements in the metrics, artifacts remained and the metrics varied and the standard deviation of the HD and ASSD were large.

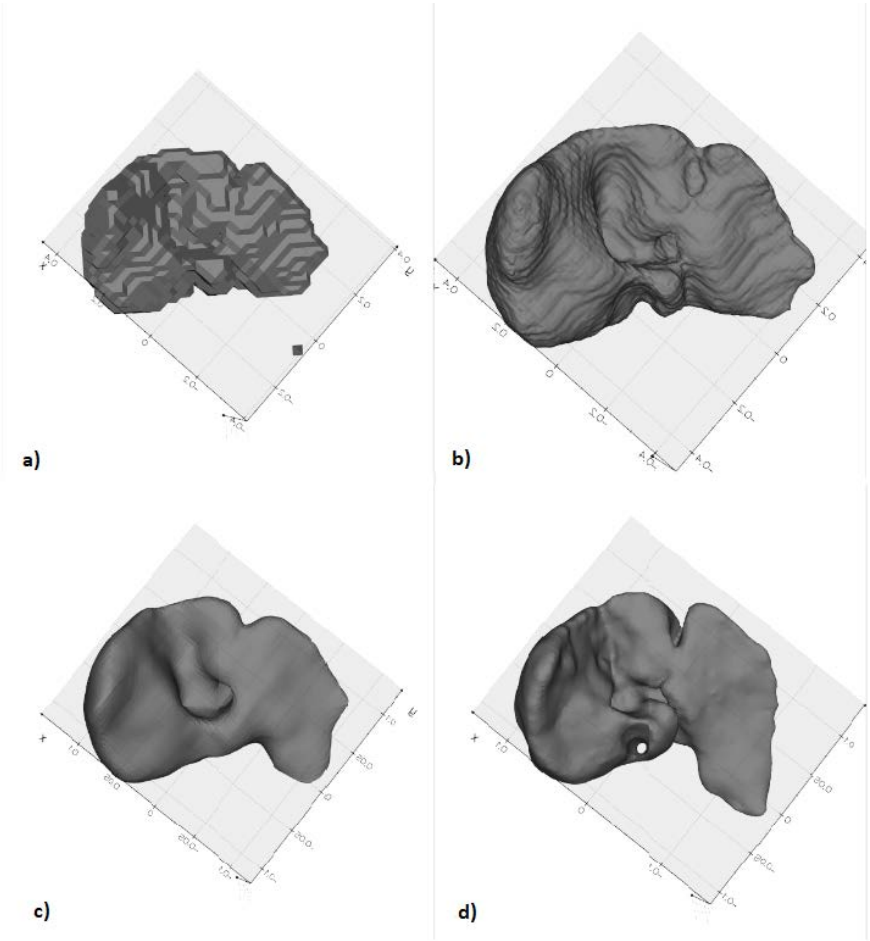


Figure 8: Visual results comparing the baseline and the proposed implicit pipeline. In the figure above, a) is 3D UNet output trained beforehand to generate voxelized segmentation results, b) is the output of the IFNet applied to a), c) is the result of the proposed end-to-end pipeline where generation of a prior voxel segmentation mask is forgone, and d) is the ground truth mesh volume for the corresponding results. As is evident from a) and b), the output of the IF-Net follows the UNet output very closely for the baseline, while the result of the end-to-end pipeline c) remains more truthful to the ground truth d) based on the presence of the structure on the lower side of the liver. Fewer artifacts are generated from the whole image is used as an input in c), compared to the visibly excessive amount of staircase effects induced in the baseline b) where the IFNet encoder performs feature extraction operations on the UNet output.

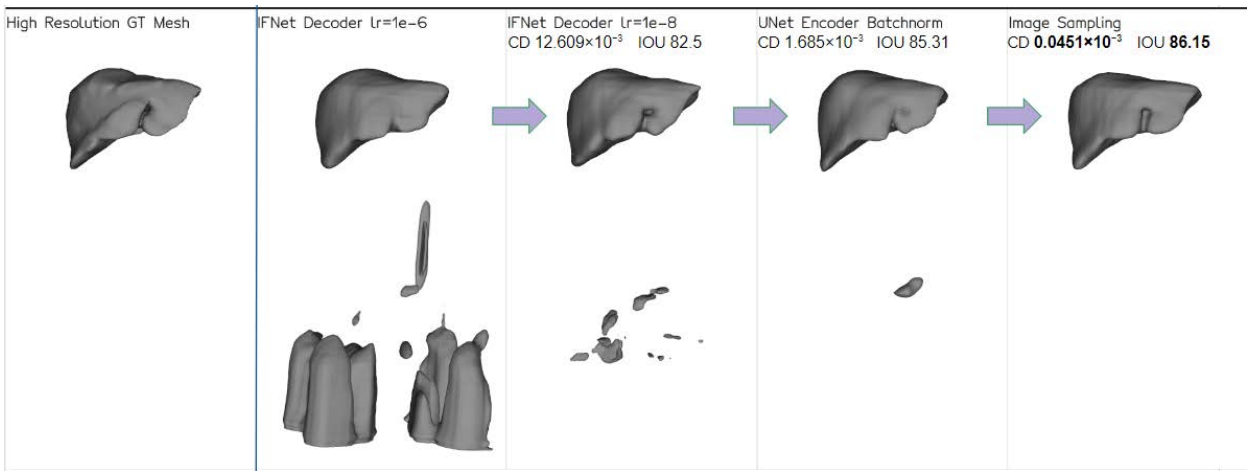


Figure 9: The resulting outputs of different optimization stages.

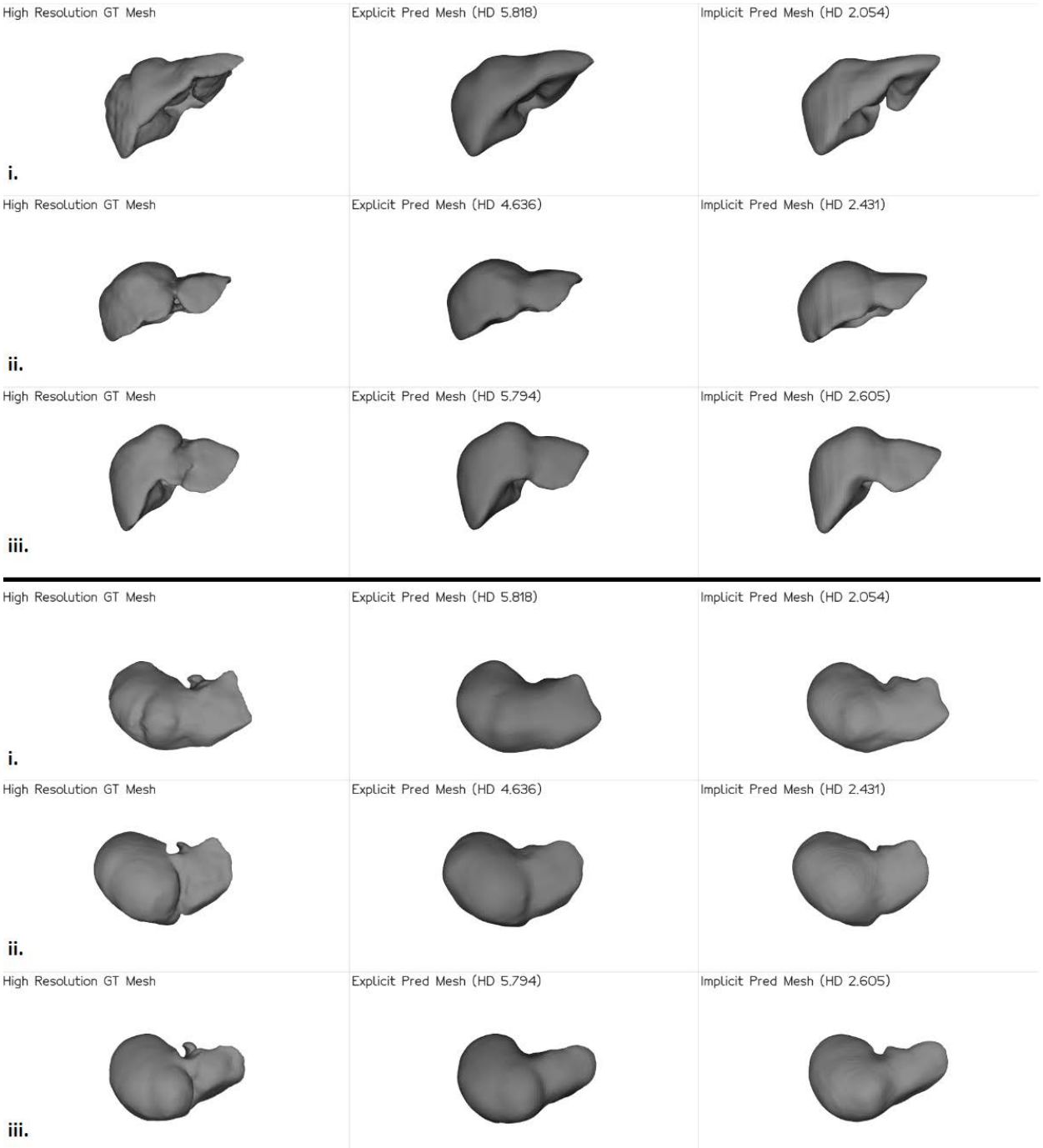


Figure 10: Different views of the ground truth mesh and explicit predicted and implicit predicted mesh of 3 images i, ii, and iii.

6. Conclusions

In this work we perform a comprehensive study to compare the benefits of explicit and implicit pipelines for 3D surface reconstruction of the liver, and further propose and optimize an implicit pipeline to output occupancy values from 3D CT images in and end-to-end manner. Our implicit pipeline with the proposed sampling outperforms the explicit and baseline implicit pipelines in all aspects, and exhibit great consistency in the surface reconstructions. Artifacts produced for difficult images by the explicit pipeline are not produced by the implicit pipeline, and training is simpler with less time being taken by the implicit pipeline, with no template initialization being required for the implicit pipeline. Further analysis of the different methods of combining features extracted from the 3D CT image using 3D CNNs with the implicit decoder will be the focus of future work, in addition to extension of the applica-

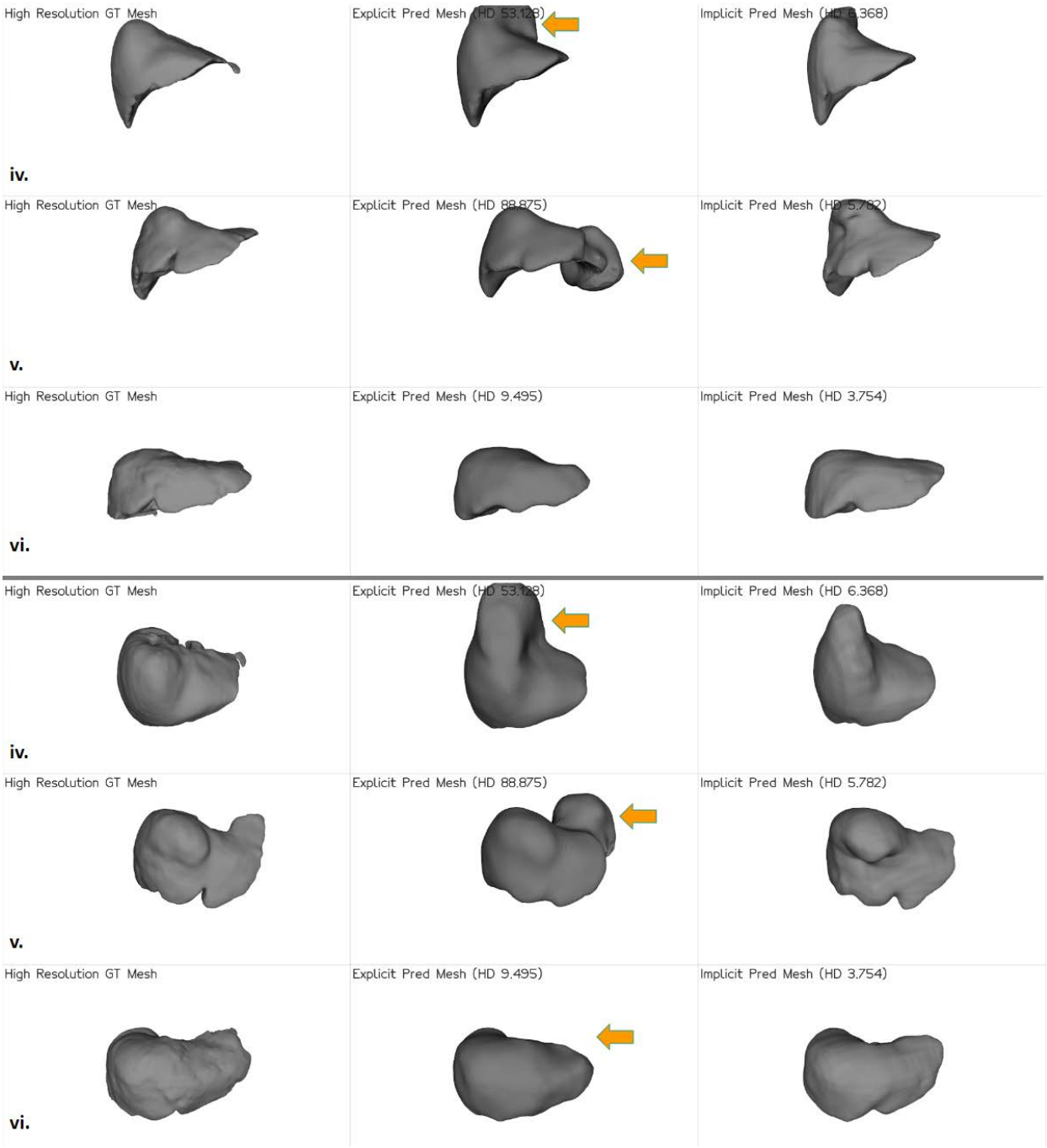


Figure 11: Different views of the ground truth mesh and explicit predicted and implicit predicted mesh of 3 images iv, v, and vi highlighting the differences for difficult images as indicated by the arrows.

tion of the pipeline to other organs.

Acknowledgments

I would like to express my gratitude to my primary supervisor, Dr. Vinkle Srivastav, for his guidance and unwavering patience throughout my thesis. I would also like to thank professor Nicolas Padoy for his support and insight into the project. Lastly, I am grateful to my classmates in the MaIA master program for their belief in my capabilities, and to my colleagues and seniors in the CAMMA research group for their support, as well my family for their encouragements from afar.

References

Alom, M.Z., Hasan, M., Yakopcic, C., Taha, T.M., Asari, V.K., 2018. Recurrent residual convolutional neural network based on

u-net (r2u-net) for medical image segmentation. arXiv preprint arXiv:1802.06955 .

Balsiger, F., Soom, Y., Scheidegger, O., Reyes, M., 2019. Learning shape representation on sparse point clouds for volumetric image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22, Springer. pp. 273–281.

Bongratz, F., Rickmann, A.M., Pölsterl, S., Wachinger, C., 2022. Vox2cortex: fast explicit reconstruction of cortical surfaces from 3d mri scans with geometric deep neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20773–20783.

Cai, J., Xia, Y., Yang, D., Xu, D., Yang, L., Roth, H., 2019. End-to-end adversarial shape learning for abdomen organ deep segmentation, in: Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10, Springer. pp. 124–132.

Catania, R., Furlan, A., Smith, A.D., Behari, J., Tublin, M.E., Borhani, A.A., 2021. Diagnostic value of mri-derived liver surface nodularity score for the non-invasive quantification of hepatic fibrosis in non-alcoholic fatty liver disease. European Radiology 31, 256–263.

Chabra, R., Lenssen, J.E., Ilg, E., Schmidt, T., Straub, J., Lovegrove, S., Newcombe, R., 2020. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16, Springer. pp. 608–625.

Chibane, J., Alldieck, T., Pons-Moll, G., 2020a. Implicit functions in feature space for 3d shape reconstruction and completion, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 6970–6981.

Chibane, J., Pons-Moll, G., et al., 2020b. Neural unsigned distance fields for implicit function learning. Advances in Neural Information Processing Systems 33, 21638–21652.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3d u-net: learning dense volumetric segmentation from sparse annotation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19, Springer. pp. 424–432.

Dangi, S., Linte, C.A., Yaniv, Z., 2019. A distance map regularized cnn for cardiac cine mr image segmentation. Medical physics 46, 5637–5651.

Elkassem, A.A., Allen, B.C., Lirette, S.T., Cox, K.L., Remer, E.M., Pickhardt, P.J., Lubner, M.G., Sirlin, C.B., Dondlinger, T., Schmainda, M., et al., 2022. Multiinstitutional evaluation of the liver surface nodularity score on ct for staging liver fibrosis and predicting liver-related events in patients with hepatitis c. American Journal of Roentgenology 218, 833–845.

Fu, K., Peng, J., He, Q., Zhang, H., 2021. Single image 3d object reconstruction based on deep learning: A review. Multimedia Tools and Applications 80, 463–498.

Ginès, P., Krag, A., Abraldes, J.G., Solà, E., Fabrellas, N., Kamath, P.S., 2021. Liver cirrhosis. The Lancet 398, 1359–1376.

Han, X.F., Laga, H., Bennamoun, M., 2019. Image-based 3d object reconstruction: State-of-the-art and trends in the deep learning era. IEEE transactions on pattern analysis and machine intelligence 43, 1578–1604.

Hoppe, H., DeRose, T., Duchamp, T., McDonald, J., Stuetzle, W., 1993. Mesh optimization, in: Proceedings of the 20th annual conference on Computer graphics and interactive techniques, pp. 19–26.

Kavur, A.E., Gezer, N.S., Barış, M., Şahin, Y., Özkan, S., Baydar, B., Yüksel, U., Kılıkçıer, Ç., Olut, Ş., Akar, G.B., et al., 2020. Comparison of semi-automatic and deep learning-based automatic methods for liver segmentation in living liver transplant donors. Diagnostic and Interventional Radiology 26, 11.

Khan, M.O., Fang, Y., 2022. Implicit neural representations for medical imaging segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V, Springer. pp. 433–443.

Kim, S.W., Kim, Y.R., Choi, K.H., Cho, E.Y., Song, J.S., Kim, J.E., Kim, T.H., Lee, Y.H., Yoon, K.H., 2020. Staging of liver fibrosis by means of semiautomatic measurement of liver surface nodularity in mri. American Journal of Roentgenology 215, 624–630.

Kong, F., Wilson, N., Shadden, S., 2021. A deep-learning approach for direct whole-heart mesh reconstruction. Medical image analysis 74, 102222.

Kotowski, K., Kucharski, D., Machura, B., Adamski, S., Becker, B.G., Krason, A., Zarudzki, L., Tessier, J., Nalepa, J., 2023. Detecting liver cirrhosis in computed tomography scans using clinically-inspired and radiomic features. Computers in Biology and Medicine 152, 106378.

Liao, Y., Donne, S., Geiger, A., 2018. Deep marching cubes: Learning explicit surface representations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2916–2925.

Liu, Y., Duan, Y., Zeng, T., 2022. Learning multi-level structural information for small organ segmentation. Signal Processing 193, 108418.

Lorensen, W.E., Cline, H.E., 1987. Marching cubes: A high resolution 3d surface construction algorithm. ACM siggraph computer graphics 21, 163–169.

Marimont, S.N., Tarroni, G., 2022. Implicit u-net for volumetric medical image segmentation, in: Medical Image Understanding and Analysis: 26th Annual Conference, MIUA 2022, Cambridge, UK, July 27–29, 2022, Proceedings, Springer. pp. 387–397.

Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A., 2019. Occupancy networks: Learning 3d reconstruction in function space, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4460–4470.

Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R., 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM 65, 99–106.

Navarro, F., Shit, S., Ezhov, I., Paetzold, J., Gafita, A., Peeken, J.C., Combs, S.E., Menze, B.H., 2019. Shape-aware complementary-task learning for multi-organ segmentation, in: Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10, Springer. pp. 620–627.

Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S., 2019. Deepsdf: Learning continuous signed distance functions for shape representation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 165–174.

Perez, A.A., Noe-Kim, V., Lubner, M.G., Graffy, P.M., Garrett, J.W., Elton, D.C., Summers, R.M., Pickhardt, P.J., 2022. Deep learning ct-based quantitative visualization tool for liver volume estimation: defining normal and hepatomegaly. Radiology 302, 336–342.

Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 652–660.

Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H., 2019. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 2304–2314.

Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G., 2020. Implicit neural representations with periodic activation functions. Advances in Neural Information Processing Systems 33, 7462–7473.

Smith, A.D., Branch, C.R., Zand, K., Subramony, C., Zhang, H., Thaggard, K., Hosch, R., Bryan, J., Vasanji, A., Griswold, M., et al., 2016. Liver surface nodularity quantification from routine ct images as a biomarker for detection and evaluation of cirrhosis. Radiology 280, 771–781.

Tewari, A., Fried, O., Thies, J., Sitzmann, V., Lombardi, S., Sunkavalli, K., Martin-Brualla, R., Simon, T., Saragih, J., Nießner, M., et al., 2020. State of the art on neural rendering, in: Computer

Graphics Forum, Wiley Online Library. pp. 701–727.

Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.G., 2018. Pixel2mesh: Generating 3d mesh models from single rgb images, in: Proceedings of the European conference on computer vision (ECCV), pp. 52–67.

Wasserthal, J., Meyer, M., Breit, H.C., Cyriac, J., Yang, S., Segeroth, M., 2022. Totalsegmentator: robust segmentation of 104 anatomical structures in ct images. arXiv preprint arXiv:2208.05868 .

Wickramasinghe, U., Remelli, E., Knott, G., Fua, P., 2020. Voxel2mesh: 3d mesh model generation from volumetric data, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23, Springer. pp. 299–308.

Xue, Y., Tang, H., Qiao, Z., Gong, G., Yin, Y., Qian, Z., Huang, C., Fan, W., Huang, X., 2020. Shape-aware organ segmentation by predicting signed distance maps, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 12565–12572.

Yao, L., Jiang, P., Xue, Z., Zhan, Y., Wu, D., Zhang, L., Wang, Q., Shi, F., Shen, D., 2020. Graph convolutional network based point cloud for head and neck vessel labeling, in: Machine Learning in Medical Imaging: 11th International Workshop, MLMI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings 11, Springer. pp. 474–483.

Ye, M., Huang, Q., Yang, D., Wu, P., Yi, J., Axel, L., Metaxas, D., 2021. Pc-u net: Learning to jointly reconstruct and segment the cardiac walls in 3d from ct data, in: Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers 11, Springer. pp. 117–126.