MAIA

ERASMUS MUNDUS

JOINT MASTER
IN MEDICAL IMAGING
AND APPLICATIONS

# Joint Master in Medical Imaging and Applications

# Master Thesis Proceedings

**Promotion 2022-24**

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

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; 2024).

# 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.

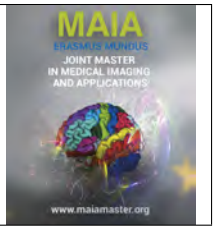
We finally want to thank and congratulate all the students for their effort done during this last semester of the Joint Master Degree in Medical Imaging and Applications.

MAIA Master Academic and Administrative Board

# Contents

Medical Imaging and Applications

Master Thesis, June 2024




# Mitigating Catastrophic Forgetting in Multiple Sclerosis Lesion Segmentation using Elastic Weight Consolidation

Luisana Álvarez[a,b], Sergi Valverde[a], Xavier Lladó[b]

[a]*Tensor Medical, Girona, Spain*
[b]*Vicorob Institute, University of Girona, Girona, Spain*

**Abstract**

Multiple sclerosis (MS) is a chronic autoimmune disease that represents the most common cause of non-traumatic disability in young adults. Accurate detection of MS lesions from magnetic resonance imaging plays a crucial role in clinical practice, from initial diagnosis to disease prognosis and treatment evaluation. Recent advancements in deep learning-based automatic MS lesion segmentation have shown promising results. However, these models often suffer from limited generalizability when applied to data with domain shifts, such as variations in image acquisition protocols, scanner, contrast, noise level or magnetic field strength. Transfer learning techniques offer a potential solution by leveraging knowledge from a source domain to adapt the model to a new target domain. While being successful in achieving target domain adaptation, transfer learning can lead to catastrophic forgetting, resulting in a significant performance drop on the source domain. Continuous learning aims to address this issue by enabling the model to retain knowledge from previous domains while adapting to new ones.

This work investigated the potential of continuous learning techniques, more specifically focusing on Elastic Weight Consolidation (EWC), to mitigate catastrophic forgetting in the context of domain-incremental learning for MS lesion segmentation. It further explored the application of EWC in a few-shot domain adaptation setting to analyze the potential for reducing the number of target domain images required for successful adaptation while preserving source domain knowledge. To evaluate the proposed approaches, several public international and in-house datasets were employed. The results demonstrated the effectiveness of EWC in mitigating catastrophic forgetting in both full-training and few-shot scenarios, enabling proper adaptation to the target domain. Additionally, EWC eliminated the need for source domain images during target domain training, addressing storage requirements and potential privacy concerns associated with medical data.

*Keywords:* Multiple sclerosis, lesion segmentation, continuous learning, transfer learning, catastrophic forgetting

## 1. Introduction

Multiple sclerosis (MS) is a chronic autoimmune and degenerative disease that affects the central nervous system, characterized by areas of inflammation, demyelination and chronic clerotic plaques (lesions), mainly in the white matter tissue (Compston and Coles, 2008). It is the most common cause of non-traumatic disability among young adults. MS is a very prevalent disease, reaching 2.9 million diagnosed patients in 2023, which means a prevalence of 36 per 100.000 people around the world. It is more common among women (69%) than among men (31%), and several studies have shown a higher prevalence in regions further from the equator, probably linked to lower vitamin D levels (Multiple Sclerosis International Federation, 2021).

Magnetic resonance imaging (MRI) is one of the main tools for diagnosis and follow-up of MS. MRI scans can provide quantitative information such as the number and volume of lesions, brain atrophy, as well as the appearance of new lesions in follow-up scans of the same patient. This quantitative analysis allows the assessment of disease progression and evaluation of therapies (Lladó et al., 2012). McDonald criteria (Filippi et al., 2022) states that MRI scans should show evidence of damage in at least two separate areas of the central nervous system, including brain, spinal cord and

optic nerves (dissemination in space) and at different points in time (dissemination in time) in order to diagnose a patient with MS. Several MRI sequences are used in the context of MS, such as T2-weighted (T2w), T1-weighted (T1w) or fluid-attenuated inversion recovery (FLAIR), sometimes together with Gadolinium enhancement. As illustrated in Figure 1, MS lesions can be seen as areas of low signal intensity in T1w images or hyperintensity areas in T2w or FLAIR images. T2w sequences can be used to detect MS lesions, but they present the major drawback of having similar intensities between lesions and cerebrospinal fluid. On the other hand, FLAIR images provide better discrimination between lesions and healthy tissues, while T1w scans provide the best contrast between different tissues.

Since a quantitative analysis of the lesions is key to assess the progression of the disease and evaluate different treatment options, segmentation of lesions becomes an important tool in clinical practice. This can be achieved manually, but this is not only time consuming and tedious, but it is also affected by intra-observer and inter-observer variability (Zeng et al., 2020). On the other hand, automatic segmentation is not trivial. Some of the main challenges are the changes in shape, location and volume of lesions across patients, the presence of artifacts or the low resolution of MRI scans, the intensity distribution overlap between healthy tissue and lesions or the high imbalance between the volume of the plaques and healthy tissue.

Many different automatic segmentation methods have been developed, and the focus on the recent years has been on deep learning approaches, in particular, Convolutional Neural Networks (CNNs). To further advance the field and objectively compare different techniques, several public challenges have been organized, providing benchmark datasets and evaluation metrics. These challenges include the MICCAI 2016 MS Lesion Segmentation Challenge (Commowick et al., 2018), the MICCAI 2021 Challenge focusing on newly appearing lesions (Commowick et al., 2021), the White Matter Hyperintensity MICCAI 2017 Challenge (Kuijf et al., 2019) and the Shifts Challenge 2023 on MS lesion segmentation simulating domain shifts scenarios (Malinin et al., 2022).

One of the main drawbacks of deep learning models is their lack of adaptability (Pianykh et al., 2020) when tested on data that differs from the one they were trained on (see Figure 2 (a)). This phenomenon, known as domain shift, occurs when the statistical distribution of the inference data (data the model is applied to) differs from the source data (data the model was trained on) (Guan and Liu, 2022). This becomes very important in medical imaging, where variations in image acquisition, scanner, contrast, noise level, magnetic field strength (1.5T vs. 3T) or presence of bias field (intensity inhomogeneity) usually lead to poor generalization capabilities (Valverde et al., 2019).

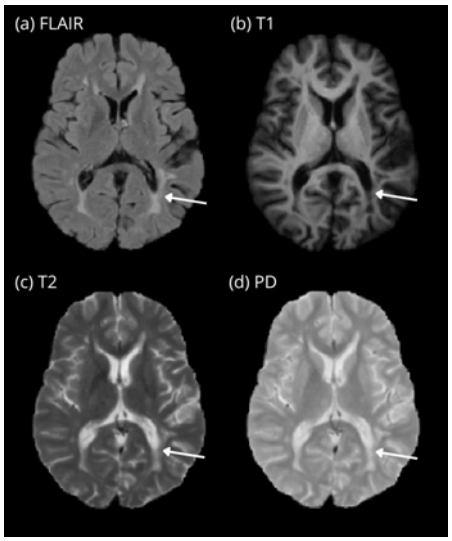



Figure 1: Variations in MS lesion appearance across MRI modalities. Arrows point to example lesions to illustrate the differences in contrast. MRI scans of an MS patient in four modalities: (a) FLAIR, (b) T1-weighted (T1w), (c) T2-weighted (T2w), and (d) Proton Density (PD). Data source: Shifts Challenge 2023 (Malinin et al., 2022).

Transfer learning (TL) emerges as a possible solution to the above mentioned problem, by using the knowledge gained in solving a specific problem to improve the performance on a target task with a different underlying data distribution (Karimi et al., 2021). It is widely adopted in the medical domain to adapt a model to a new dataset, specially if there is limited available data. However, the main focus of TL is to leverage prior knowledge, rather than retaining it, leading to an abrupt loss in performance in the source dataset once the model is retrained on the target dataset (Pianykh et al., 2020). This phenomenon is known as catastrophic forgetting (see Figure 2 (b)). Continuous learning (CL) arises as a solution to this issue, with the objective of retaining knowledge from previous tasks while adapting to new tasks. Therefore, it can be stated that, while TL only focuses on the target domain, CL focuses both on the source and target domains (Kumari et al., 2024). In essence, CL aims to continuously expand the model's capacity in an incremental way, allowing it to learn and integrate new information without forgetting past knowledge (see Figure 2 (c)). CL comes with different variants, such as task incremental learning (Baweja et al. 2018, Kaustaban et al. 2022), class incremental learning (Ozdemir et al. 2018, Liu et al. 2022) or domain incremental learning (Karani et al. 2018, van Garderen et al. 2019). The ability to continuously learn and solve new tasks without catastrophic forgetting makes CL a highly desirable approach in various deep learning applications. Given its potential to overcome limitations of current models and enable real-world applications that require knowledge preservation, CL is currently a hot topic in the deep learning research field.

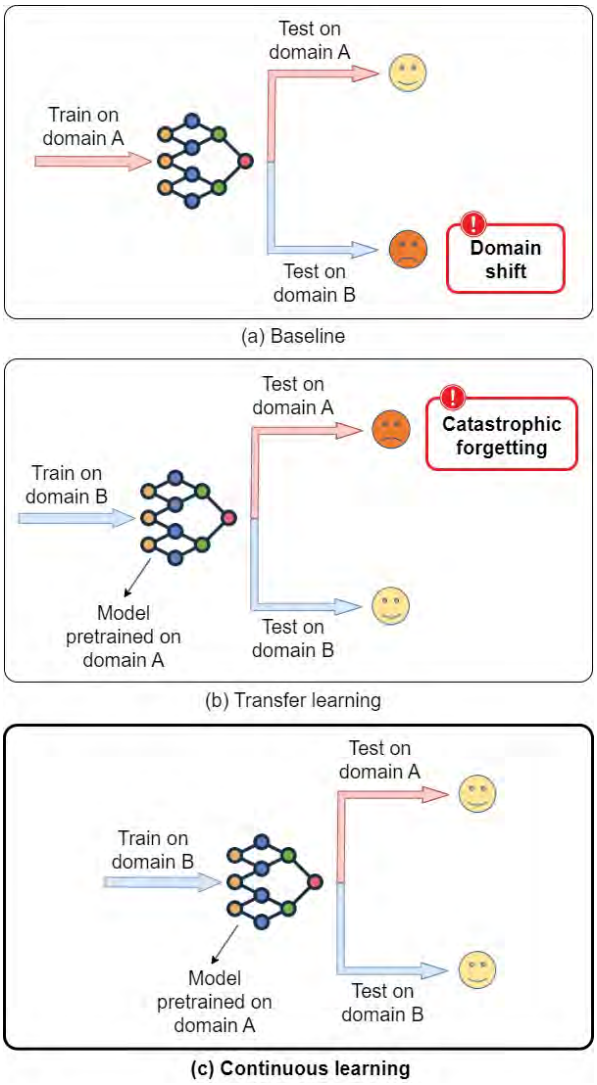



Figure 2: Schematic representation of the problems faced in this work. (a) Baseline models fail to generalize to unseen data with different distributions (domain shift) . (b) TL effectively adapts the model to the target domain, but at the cost of catastrophic forgetting of the source domain. (c) CL aims to mitigate catastrophic forgetting while still allowing the model to adapt to the target domain.

### *1.1. Objectives and contributions*

The main objective of this work is to study the domain shift problem in MS lesion segmentation using deep learning models, incorporating CL strategies to tackle the catastrophic forgetting issue present in deep learning models and TL strategies. To this end, a deep learning model using 3D patches is developed for MS lesion segmentation. A simple U-Net architecture is chosen in order to clearly understand the impact of TL and CL techniques and to avoid architectural complexities influencing this evaluation. The performance of this baseline model is evaluated on in-domain and out-domain images from both public (White Matter Hyperintensity Challenge 2017 (Kuijf et al., 2019) and Shifts Challenge 2023 (Malinin et al., 2022)) and in-house (Vall d'Hebron Hospital in Barcelona, Spain) datasets, to assess the severity of the domain shift problem.

To improve generalizability, TL is employed as a domain adaptation technique. Different experiments are conducted to understand the effect of unfreezing different sections of the encoder and the encoder of network. Moreover, one-shot and few-shot domain adaptation approaches are explored, due to their potential to improve generalizability with minimal data from the target domain. This is particularly advantageous in medical imaging where acquiring large amounts of labeled data can be expensive and time-consuming.

Finally, in order to alleviate catastrophic forgetting in the source domain, different regularization-based CL approaches are studied. These techniques show successful results in preserving previously acquired knowledge while effectively learning on the new domain, in both full-training and few-shot domain adaptation scenarios. Notably, this is achieved without requiring the source domain data to be present during training. This overcomes a major limitation of traditional adaptation techniques, which often need access to the original data for retraining.

## 2. Related work

### *2.1. MS lesions segmentation*

The first proposals for MS lesion segmentation relied on traditional techniques. Some of them were based on segmenting brain tissues and detected lesions as outliers that were not well explained by a statistical model such as expectation-maximization (Van Leemput et al. 2001, Roura et al. 2015). Some other approaches, were based on directly detecting lesions based on their properties in the images, for example, by training a classifier like K-Nearest Neighbors using the image intensities and manual segmentations (Petronella et al., 2005). However, the emergence of deep learning has significantly advanced the field of MS lesions segmentation, thanks to the ability of these models to learn complex patterns without the need of explicit feature engineering. The superior performance and robustness of deep learning methods compared to traditional approaches have led to a significant shift in research focus towards deep learning for MS lesion segmentation.

Early attempts at deep learning for MS lesions segmentation often relied on architectures composed by a series of convolutional layers for feature extraction followed by fully-connected layers for pixel-based classification (Valverde et al. 2017). Even though these approaches achieved successful results, they tended to lose spatial context. Nowadays, to address this limitation, the main focus of research is on U-Net-shaped architectures, that incorporate skip connections linking the feature maps from the encoder to the decoder in order to preserve spatial information. Due to the success of U-Net models in this task (Ronneberger et al.,

2015), most of the recent work has been focused on variations of this architecture, such as the incorporation of attention mechanisms (hu et al. 2020, Hashemi et al. 2022, Gamal et al. 2023), the creation of multiple branches to handle feature extraction separately for each modality (Aslani et al., 2019), the substitution of the U-Net encoder with a pre-trained network like VGG19 (Krishnamoorthy et al., 2022) or the addition of pre-activation residual blocks (Ashtari et al., 2022). Some other authors opted for keeping the U-Net original architecture but proposing novel techniques like a lesion-specific loss function (Zhang et al., 2021) or sequence dropout (Feng et al., 2019). Finally, the nnU-Net framework (Isensee et al., 2021) has emerged as a standard for medical image segmentation challenges due to its ability to achieve high performance with minimal user intervention. nnU-Net leveraged the U-Net architecture's strengths by incorporating self-adaptation capabilities, allowing the model to automatically configure its parameters based on the specific input dataset, reducing the need for extensive manual tuning.

Regarding how data is fed into network, the general tendency in MS lesions segmentation is to work with patches instead of full volumes. Authors have proposed different strategies, including 2D (Krishnamoorthy et al., 2022), 2.5D (Zhang et al., 2019) and 3D approaches (Valverde et al., 2017). The current trend is to employ 3D approaches, due to their ability to capture spatial context more effectively. Furthermore, the selection of the imaging modality is also diverse. FLAIR is the principal modality used nowadays for automatic MS segmentation, becoming a standard in MS imaging protocols. This is due to the good contrast this modality offers between lesions and healthy brain tissue. Even though it can be used together with other modalities such as T1w, T2w, or PD, recent automatic approaches have been mainly focused on using only FLAIR for automatic segmentation.

### *2.2. Transfer learning for domain adaptation*

TL is widely used in the medical domain in order to take advantage of pretrained networks in cases of limited labeled data, a common scenario in medical imaging segmentation applications. Ghafoorian et al. (2017) analysed the effect of the training set size and the number of unfreezed layers when performing TL for domain adaptation applied to MS lesion segmentation.Valverde et al. (2019) also studied the impact of the amount of unfreezed parameters but in one-shot domain adaptation scenarios, assessing how did the lesion load of the chosen subject impact in the TL results. Unsupervised approaches have also been proposed for MS lesions segmentation. This is the case of the work of Kushibar et al. (2021) in which a transductive TL approach was proposed, aiming to align the feature distribution of the source and target domains. The convolutional and fully connected layers were forced to produce similar activation maps by minimising the histogram distribution differences.

Nevertheless, what all the previous works had in common was that they employed architectures with convolutional layers for feature extraction followed by fully connected layers rather than U-Net shaped networks. Existing literature investigating optimal approaches for TL with U-Net shaped architectures is limited, particularly regarding the selection of layers to unfreeze during training. Shirokikh et al. (2020) conducted different experiments with the U-Net architecture, comparing the results of unfreezing layers on the encoder, the decoder and the full network for domain adaptation in brain structures segmentation. They concluded that encoder layers contain more domain specific information than decoder layers, being a best choice in domain adaptation problems. However, this area of research has not been extensively studied and represents an opportunity for further exploration in the field of MS lesion segmentation.

### *2.3. Continuous learning*

CL scenarios can be categorized according to the differences between the source and target datasets (Kumari et al., 2024):

- **Data-incremental** or **instance incremental scenario** if the data comes from the same data distribution. It is the least challenging among all scenarios (Ravishankar et al. 2019, Kaustaban et al. 2022).
- **Class-incremental scenario** if the goal is to adapt the model to incorporate new classes (Ozdemir et al. 2018, Liu et al. 2022).
- **Task-incremental scenario** if each episode has a disjoint label space, meaning it would be evaluated only on the current episode data (Baweja et al. 2018, Kaustaban et al. 2022).
- **Domain-incremental scenario** if the shifts of the data are due to different domains (different imaging modality, acquisition protocols, scanners, contrast agents...). It is one of the most common scenarios in medical imaging (Karani et al. 2018, van Garderen et al. 2019).

Moreover, the are several different CL strategies to prevent catastrophic forgetting when learning new tasks:

- **Rehearsal-based approaches** store previous tasks' data in a small memory buffer to be used while training on new tasks. The stored data can be the original images (experience-replay based) (Perkonigg et al., 2021), deep features (latent replay-based) (Srivastava et al., 2021) or generated pseudo samples (generative replay-based) (Li

et al., 2023) and can be selected via different heuristics. These kind of approaches can violate privacy concerns, specially when storing the original images, which is a problem when dealing with medical data, and usually have high memory requirements.

- **Regularization-based methods** aim to control weight update within the training of the model to minimize forgetting the previous learning, either through knowledge distillation from a teacher model to a student model (data-focused regularization) (Li and Hoiem, 2018) or by penalizing large changes on important parameters for previous tasks (prior-focused regularization) (van Garderen et al., 2019).
- **Architectural-based methods** assign to each task a set of parameters, either by fixing the architecture (limited by the network's capacity) (Bayasi et al., 2021) or dynamically extending the network (increasing memory requirements with each new task) (Yan et al., 2021).

Depending on the level of supervision, CL methods can be classified into task aware or task agnostic depending on if they require or not information about which task the samples belong to. Moreover, new tasks can be introduced in a rigid from (abrupt change from one episode to another) or in a non-rigid way, by interleaving samples of the previous and current task or by gradually increasing the amount of new data while decreasing the amount of old data.

From now on, the focus will be on regularization-based CL methods. This choice was motivated by several key advantages. Firstly, unlike rehearsal-based methods, they do not require storing data from previous tasks during training on new tasks. This eliminates potential privacy concerns associated storing with medical images and reduces memory requirements. Secondly, they avoid the substantial memory overhead incurred by some architectural-based methods that dynamically expand the network for each new task. Finally, they offer more flexibility compared to strictly assigning network parts to each task in other architectural approaches. Within the realm of regularization-based methods, a particular focus is placed on prior-focused approaches, since data-focused methods require an additional model for the new task (student model) to perform knowledge distillation from the previous task's model (teacher model). Prior-focused methods, on the other hand, leverage information about the model's previous state directly, making them more lightweight and efficient.

One of the most well-known prior-based regularization methods is Elastic Weight Consolidation (EWC), proposed by Kirkpatrick et al. (2017) for supervised image classification tasks and reinforcement learning scenarios for video games. This method aimed to emulate synaptic consolidation of human brains to reduce catastrophic forgetting by adding a penalty term in the loss function that slowed down learning on specific weights that were important for previous tasks. The importance of each weight was computed based on an approximation of the Fisher information matrix of the outputs of the network for the previous tasks data. EWC has been adapted to different medical tasks. For instance, van Garderen et al. (2019) applied it for a glioma segmentation problem, in order to perform TL from a public dataset containing low and high-grade glioma to an in-house dataset containing non-enhancing low-grade glioma. Another example is the work of Baweja et al. (2018), who used EWC to learn sequentially two different tasks: first, brain tissue segmentation and then, white matter lesions segmentation. Finally, Chen and Tang (2022) developed a CL pipeline with EWC penalty for breast tumor classification in two different scenarios (class-incremental and instance-incremental).

Memory Aware Synapses (MAS) was proposed by Aljundi et al. (2018) as a regularization-based CL method, and was also based on the idea of keeping important parameters learnt for previous tasks. It differed from EWC in how the importance of the parameters was computed. In this case, it was based on the gradients of the squared $l_2$ norm of the learned function output. This approach was adapted for domain-incremental CL for brain segmentation by Özgün et al. (2020), who also proposed an alternative regularization approach, employing the importance of parameters to define parameter-specific learning rate to protect performance on previous tasks, instead of applying the regularization as an extra loss term. Moreover, they proposed a pruning strategy by freezing important parameters during training on new tasks.

Zhang et al. (2023) developed a regularization method that, as in EWC and MAS strategies, penalized changes on parameters that were important for previous tasks. However, this method differentiated itself by focusing on parameters sensitive to shape and semantic features, aiming to specifically retain such knowledge from past tasks. Synaptic Intelligence (Zenke et al., 2017) reduced catastrophic forgetting by identifying connections that were not strongly tied to previous tasks so, when new tasks arrived, the network focused on adapting the weights of these uncommitted synapses. The main drawback of this method is that it requires extra parameters to label committed synapses. Finally, Distributed Weight Consolidation (McClure et al., 2018) allowed training different networks for each task, in this case brain segmentation datasets from different sites, and later consolidate those weights on a single network.

While a review of state-of-the-art CL methods revealed promising techniques for alleviating catastrophic forgetting in general, its application to domain-incremental MS lesion segmentation remains an under-

Table 1: Continuous learning for medical imaging state-of-the-art overview.

| CL strategy | | Reference | CL scenario | Task description |
|---|---|---|---|---|
| Rehearsal-based | Latent replay | Ravishankar et al. (2019) | Data-incremental | X-ray pneumothorax classification |
| | | Srivastava et al. (2021) | Domain-incremental | Chest X-ray classification |
| | | Liu et al. (2022) | Task-incremental | CT multi-organ segmentation |
| | | Karthik et al. (2022) | Domain-incremental | MS lesion segmentation |
| | Experience replay | Perkonigg et al. (2021) | Domain-incremental | Cardiac segmentation and lung nodule detection |
| | Generative replay | Li et al. (2023) | Domain-incremental | Cardiac image segmentation |
| Regularization-based | Data-focused | Ozdemir et al. (2018) | Class-incremental | Segmentation of different structures |
| | Prior-focused | Baweja et al. (2018) | Task-incremental | From brain segmentation to white matter lesions segmentation |
| | | van Garderen et al. (2019) | Domain-incremental | Glioma segmentation |
| | | Özgün et al. (2020) | Domain-incremental | Brain structure segmentation |
| | | Chen and Tang (2022) | Class and data-incremental | Breast tumor tissue classification |
| | | Zhang et al. (2023) | Domain-incremental | Prostate and optic cup and disk segmentation |
| Architectural-based | | Karani et al. (2018) | Domain-incremental | Brain structure segmentation |
| | | McClure et al. (2018) | Domain-incremental | Brain structure segmentation |
| | | Bayasi et al. (2021) | Domain-incremental | Skin lesion classification |
| Regularization and rehearsal-based | | Kaustaban et al. (2022) | Data and task-incremental | Tumor hystopathology classification |

explored area. Karthik et al. (2022) were the first to apply CL in this specific scenario, through a rehearsal-based approach. Data from previous tasks was stored in a memory buffer and interleaved with the current domain data. The main disadvantage of this method is the necessity of having data from previous tasks available for training which, as mentioned before, not only leads to high storage requirements but also can provoke privacy violation issues.

A concise overview of the related works in CL is provided in Table 1. Motivated by the gap in research on applying CL to domain-incremental learning for MS lesion segmentation, this work aimed to address the limitations of TL by studying CL techniques that can effectively mitigate catastrophic forgetting during domain adaptation for MS lesion segmentation. The study specifically focused on CL methods that meet the following requirements, which are of special interest for companies that work on developing deep learning systems for medical image analysis:

- The method should not require images from the previous domain when training on the new domain.
- The approach should be memory-efficient, avoiding the addition of extra network parameters, training of separate networks, or storing images from the previous domain.

Considering these requirements, prior-focused regularization-based methods such as EWC or MAS emerged as suitable CL strategies for this investigation. Furthermore, the study explored the application of these CL techniques in the context of few-shot domain adaptation to analyze the potential for significant reduction in the number of target domain images required for successful adaptation.

## 3. Material and methods

### *3.1. Datasets*

#### *3.1.1. White Matter Hyperintensity MICCAI challenge 2017 (WMH2017)*

The WMH2017 dataset contained multimodal 3D brain MRI scans from 60 subjects acquired from five scanners from three different vendors (Simenes, Philips and General Electric) in three hospitals in the Netherlands and Singapore, as it can be seen in Table 2 (Kuijf et al., 2019). For each subject, T1w and FLAIR modalities were provided. For the T1w images, the dataset contained the original 3D T1w image, with the face removed, and also the 3D T1w scan registered to the FLAIR image. For both the FLAIR and the two T1w volumes, the bias field corrected versions were also included, processed using SPM12 software (Functional Imaging Laboratory, 2014). Moreover, the dataset contained a manual reference standard, consisting on manually segmented white matter lesions according to the STandards for ReportIng Vascular changes on nEuroimaging (STRIVE), made by four expert observers, in the space of the FLAIR image.

In this case, even though the organization provided a pre-processed version of the images, it was preferred to use the original images and process them as needed. The pre-processing steps included affine registration to the MNI $1 \times 1 \times 1$ mm template, skull stripping using HD-BET (Isensee et al., 2019) and bias field correction using the N4 algorithm (Tustison et al., 2010). Examples of each pre-processing step can be seen in Figure 3.

For this work, the images of the 60 patients were split into 41 images for training, 7 for validation and 12 for testing, ensuring that in each split there was approximately the same number of images from each scanner.

Table 2: Description of the datasets used for MS lesion segmentation. Scanner manufacturer and location are provided. Resolution is indicated in mm³ . Train, Val., Tes., and Total columns represent the number of samples used for training, validation, testing, and total, respectively.

| Dataset | | Location | Scanner | Resolution ($mm^3$) | Train | Val. | Test | Total |
|---|---|---|---|---|---|---|---|---|
| WMH2017 | | UMC Utretch | 3T Philips Achieva | $0.96 \times 0.95 \times 3.00$ | 12 | 4 | 4 | 20 |
| | | NUHS Singapore | 3T Siemens TrioTim | $1.00 \times 1.00 \times 3.00$ | 15 | 1 | 4 | 20 |
| | | VU Amsterdam | 3T GE Signa HDxt | $1.20 \times 1.21 \times 1.30$ | 14 | 2 | 4 | 20 |
| Shifts | MSSEG-1 | Rennes | 3T Siemens Verio | $0.50 \times 0.50 \times 1.10$ | 8 | 2 | 5 | 15 |
| | | Bordeaux | 3T GE Discovery | $0.47 \times 0.47 \times 0.90$ | 5 | 1 | 2 | 8 |
| | | Lyon | 1.5T Siemens Aera | $1.03 \times 1.03 \times 1.25$ | 10 | 2 | 17 | 29 |
| | | | 3.0 Philips Ingenia | $0.74 \times 0.74 \times 0.70$ | | | | |
| | ISBI | Best | 3T Philips Medical | $0.82 \times 0.82 \times 2.20$ | 10 | 2 | 9 | 21 |
| VH | | Vall d'Hebron Barcelona | 3T Siemens TrioTim | $0.49 \times 0.49 \times 3.00$ | 25 | 5 | 27 | 57 |

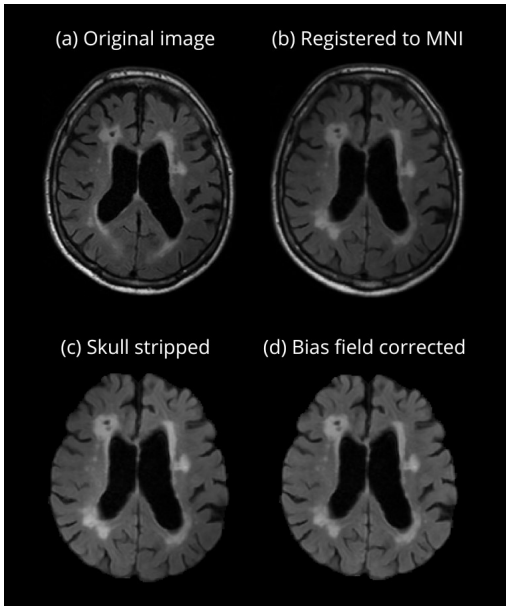



Figure 3: Pre-processing pipeline for WMH2017 dataset. (a) Original image. (b) Image after affine registration to the MNI space. (c) Image after skull stripping with HD-BET (Isensee et al., 2019). (d) Image after bias field correction using N4 algorithm (Tustison et al., 2010).

In this case, 5 different splits were chosen to perform a 5-fold cross-validation.

### *3.1.2. Shifts challenge 2023*

The Shifts dataset aimed to simulate real-world scenarios in which there is a distributional shift between training and testing or deployment data (Malinin et al., 2022). This dataset was constructed following a canonical partition, meaning that there were in-domain training, development and evaluation subsets, and also out-domain or shifted development and evaluation subsets. As it can be seen in Table 2, it combined several public datasets such as ISBI, MMSSEG-1 and PubMRI, coming from different institutions and scanners and a dataset provided by the university of Lausanne. The latter was the hidden evaluation set, thus it was not provided. Images from Rennes, Bordeaux, Lyon and Best were treated as in-domain data, while Ljiublana's correspond to the the out-domain validation set. In this work, only the in-domain data was employed, since it was preferred to use the in-house dataset as a domain-shift example (see Section 3.1.3), as there were more available images and they corresponded to real-life scans.

In this dataset[1], the images were already pre-processed, and the original images were not available. The pre-processing included denoising with non-local means (Coupe et al., 2008), skull stripping using HD-BET (Isensee et al., 2019), bias field correction with a N4 algorithm (Tustison et al., 2010) and interpolation to the 1 mm isovoxel space. In this case, as the raw images were not provided, the affine registration to the MNI $1 \times 1 \times 1$ mm template was performed afterwards.

The Shifts dataset contained T1w and FLAIR brain MRI scans from 98 different subjects. The ground-truth segmentation mask was obtained as a consensus of seven expert annotators, except for Best and Lausanne, in which only one expert rater participated. In this case, as mentioned above, the splits were already provided by the challenge organisation, containing scans from 33 subjects for training, 7 for validation and 33 for testing (all these 73 subjects were considered as in-domain samples). The remaining 25 images corresponded to the

[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

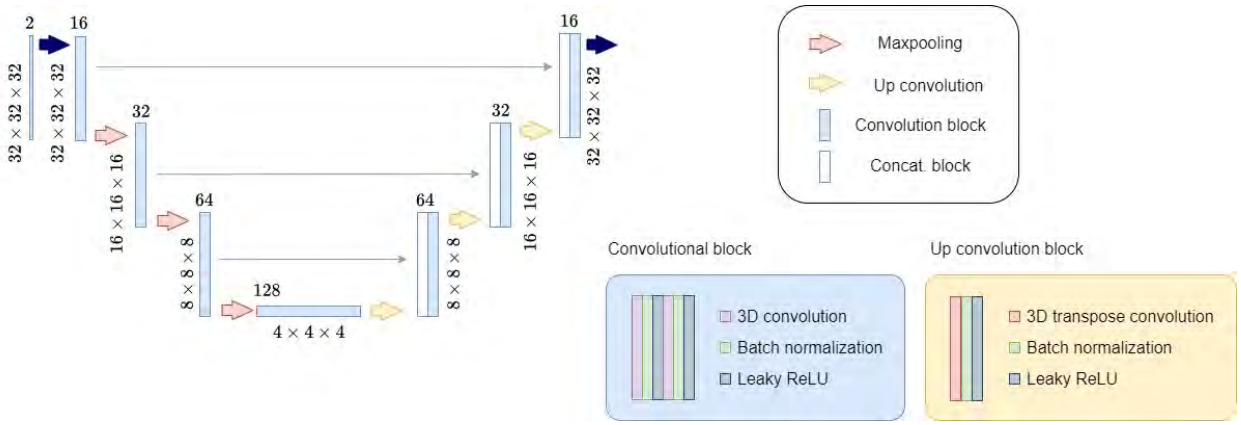



Figure 4: Schematic representation of the 3D U-Net architecture employed for MS lesion segmentation. This model served as a baseline to study TL and CL techniques.

out-domain samples that, as explained before, they were not used in this work.

#### *3.1.3. Vall d'Hebron dataset*

This in-house dataset came from the Vall d'Hebron (VH) University Hospital in Barcelona, Spain. It contained MRI scans from 57 subjects acquired from a 3T Siemens scanner. For each subject, both FLAIR and T1w modalities were provided. In this case, the images were pre-processed as stated by Valverde et al. (2019), including skull stripping, N3 bias field correction, co-registration to T1w (FSL-FLIRT) and interpolation to 1mm isovoxel space.

Since in this dataset all the images were obtained with the same scanner, they were randomly split into 25 samples for training, 5 for validation and 27 for testing (see Table 2).

### *3.2. MS lesion segmentation framework*

In this section, the proposed architecture, data manipulation and training strategies for MS lesion segmentation will be described in detail. This will be the baseline framework to later study TL and CL strategies within the domain shift problem.

#### *3.2.1. Baseline architecture*

In this work, a 3D U-Net architecture (Çiçek et al., 2016) was employed for the segmentation. While more advanced U-Net variations have been proposed in the literature, a simpler configuration was chosen in this work. This choice aimed to minimize the impact of architectural modifications on the evaluation of CL and TL methods. This allowed for a clearer understanding of how these techniques influenced the network's performance in domain-incremental learning scenarios.

The chosen model utilized an encoder-decoder structure with skip connections for efficient feature extraction and propagation. It was composed by 4 layers of convolutional blocks, with 16, 32, 64 and 128 filters per layer. In the encoder side, each convolutional block contained two sequences of convolution - batch normalization - Leaky ReLU activation, followed by a maxpooling layer for downsampling. On the other hand, each decoder block was formed by an up convolution sequence (transposed convolution - batch normalization - Leaky ReLU activation) for upsampling followed by a convolution block equal to the ones in the encoder side. A diagram on the described architecture can be seen in Figure 4. The decision of constructing each layer with two convolutions was based on experimental results, where this helped in obtaining better generalization to other domains.

#### *3.2.2. Patch sampling*

A patch-based segmentation approach was chosen for this work. The patch sampling strategy is a key feature of the segmentation pipeline. Following the patch size selection in related works by Fenneteau et al. (2021) and Salem et al. (2022), 5000 patches of $32 \times 32 \times 32$ voxels were extracted from each image. To address the problem of class imbalance, an equal number of positive and negative patches were sampled. A patch was labelled as positive or negative according to the class of its central voxel. To ensure that the selected patches provided different information from the entire anatomy of the patient, they were not randomly sampled, but uniformly extracted from both the lesions and the healthy tissue.

Patches from both FLAIR and T1w modalities were fed into the network in two separate input channels. While FLAIR MRI provided good contrast between lesions and healthy tissue, T1w sequences contributed with more structural information of the brain tissues. It is worth mentioning that, before sampling the patches, each image was normalized by subtracting its mean and dividing it by its standard deviation. For these computations, only the intensities inside the brain region were considered.

#### *3.2.3. Training strategy*

The network was trained with batches of 32 patches to balance computational efficiency with gradient update quality. To address the class imbalance problem, balanced batches were constructed, ensuring an equal number of positive and negative patches within each batch. Compared to random batch construction, balanced batches have shown empirically to improve the model's ability to learn from the minority class (lesion) and achieve better overall segmentation performance. Cross-entropy was used as the loss function and the model was trained for a maximum of 300 epochs. Alternative segmentation loss functions were explored, including DiceFocal loss and DiceCE loss. While these functions achieved performance comparable to the cross-entropy loss, they exhibited slower convergence rates. Therefore, the cross-entropy loss function was chosen and maintained throughout the study for consistency. Early stopping was applied if the validation loss did not improve in the last 20 epochs, to avoid over-fitting. Adam optimizer was selected, with an initial learning rate of $10^{-4}$ and a weight decay of $10^{-6}$. To manage the learning rate during training, a reduce-on-plateau scheduler was implemented. This scheduler automatically reduced the learning rate by a factor of 10 if the validation loss plateaued for 7 epochs, preventing the model from getting stuck in local minima. To enhance model generalization, data augmentation was implemented. This technique created variations of the original patches by applying random transformations with a 30% probability. Specifically, patches underwent random rotations within a range of -$\frac{\pi}{2}$ to $\frac{\pi}{2}$ radians, flipping along any spatial axis and affine transformations combining small rotations (-0.1 to 0.1 radians) and shearing (-0.1 to 0.1).

Three different models were trained using this strategy, one for each dataset presented in Section 3.1. These were the baseline models that were used as reference for each domain. Each model was evaluated on its corresponding testing set and on the other two datasets to assess the impact of domain shift.

For inference, a sliding window approach was employed with a 25% overlap between patches. To account for the overlap, the average probability was computed in the overlapping regions. Once the probability map for the whole volume was obtained, detected lesions smaller than 3 voxels were filtered out, as different studies have stipulated a minimum detectable diameter of 3 mm as a diagnostic criteria (Grahl et al. 2019, Filippi et al. 2019). This aimed to reduce the number of false positive lesions, at the cost of a reduction in sensitivity. The binarization threshold ($T_{bin}$) was optimized for each dataset to achieve the best trade-off between true positive and false positive lesions. This optimization involved evaluating the detection F-score for a range of threshold values from 0.1 to 0.9 with increments of 0.1. The threshold that yielded the highest F-score was then selected for each dataset.

### *3.3. Transfer learning for domain adaptation*

TL was studied as a solution for the domain shift problem. Moreover, it allowed to investigate the extent to which the model's performance on the source domain degraded after adapting to the target domain (catastrophic forgetting). The WMH2017 dataset was chosen as the source dataset due to the greater control over pre-processing steps, as opposed to the Shifts dataset, that was provided already pre-processed by the organization of the challenge. The VH dataset served as the target domain, because it represented a larger domain shift compared to the Shifts dataset, as it will be seen in Section 4.1.

To assess the impact of progressive unfreezing on adaptation, different number of layers in the encoder or decoder were progressively unfrozen and fine-tuned during training on the VH dataset. More specifically, this involved unfreezing 1 or 2 layers from the encoder or decoder and also, training with all the layers unfrozen.

To evaluate the number of images required for adaptation to the target domain, one-shot and few-shot TL approaches were also analyzed. This involved fine-tuning the model with a limited number of images (1, 2, 3, 5, or 10) from the VH dataset. The images were selected based on the lesion volume of each subject to provide as much variability as possible in this regard. Two validation images were also selected and kept fixed through all the experiments. Based on the findings from the unfreezing experiments, all one-shot and few-shot experiments were performed with the best unfreezing strategies: unfreezing all layers and unfreezing the last two encoder or decoder layers.

Due to the small number of training images in the one-shot and few-shot experiments, a different patch extraction approach was employed. Instead of a fixed number of patches, all possible patches were extracted centered on every positive voxel in the image. An equal number of negative patches was also extracted to maintain balanced training.

The hyperparameter configuration that yielded the best performance during the base model training on the WMH2017 dataset was maintained for all TL experiments. While adjustments to the learning rate towards smaller values and increased weight decay were explored to potentially reduce overfitting, the original hyperparameter settings consistently led to superior performance.

After all TL experiments, inference was performed on both the target and source domain to evaluate the model's adaptation to the new domain and assess potential catastrophic forgetting, respectively.

### *3.4. Continuous learning*

In order to study CL techniques for mitigating catastrophic forgetting during domain adaptation, different prior-focused regularization techniques were studied: Elastic Weight Consolidation (EWC), Memory Aware Synapses (MAS) and importance-based parameter-specific learning rate strategies.

#### *3.4.1. Elastic Weight Consolidation (EWC)*

EWC is a prior-focused regularization-based CL technique proposed by Kirkpatrick et al. (2017) to alleviate catastrophic forgetting. As it was explained in Section 2.3, it aims to protect parameters (weights and biases) that are important for the source domain while learning the target domain.

Training a network consists on optimizing the value of a set of parameters $\theta$ by minimizing a loss function $\mathcal{L}$. Due to the high amount of parameters in a neural network, there should be different sets of parameter values that result in the same performance. This means that there should be a solution for task B (target domain), represented by its optimal parameters $\theta^*_B$, that is close to the previous solution for task A ($\theta^*_A$) (source domain). The goal of EWC is to find this particular solution, by forcing the network to learn the task B by finding $\theta^*_B$ as close as possible to $\theta^*_A$ in the parameter space. EWC achieves this by adding a penalty to the loss function for task B $\mathcal{L}_B$. The particular thing about this penalization is that it is "elastic" (it is different for each parameter). This means that the penalty should be higher for:

- Parameters that are important for the performance in task A.
- Parameters that are getting further from the optimal values for task A.

The loss function to minimize in EWC is defined as follows:

$$\mathcal{L} = \mathcal{L}_B + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i})^2 \tag{1}$$

In this equation:

- $F_i$ represents the importance of parameter $i$.
- $\theta_i$ represents the value of the parameter $i$ in the current iteration (training on task B).
- $\theta^*_{A,i}$ represents the optimal value of the parameter $i$ for the previous task A.
- $\lambda$ controls the weight given to the old task A compared to the new one B. A higher value of $\lambda$ means a stronger emphasis on preserving A's knowledge. However, very high values can lead to the network not learning the new task, so it is important to find a trade-off.

As it can be seen in the Equation 1, this new term in the loss forces the network to minimize the difference between the current parameters and the optimal ones for the previous task A ($\theta_i - \theta^*_{A,i}$) weighted by the importance of each parameter $F_i$. Finally, the penalization term is computed as the sum of the penalization terms of each parameter.

The key component here is the parameter importance $F$. It is calculated as the diagonal of the Fisher information matrix, which measures the amount of information about a parameter $\theta_i$ that is provided by a data sample $X_j$ (in this case, a patch). It is computed as the average of the squared $1^{st}$ derivatives of the log-likelihood function with respect to the parameters. This means that the slope of the likelihood function at the true parameter $\theta$ is a measure of the amount of information provided by the observed data regarding the parameter $\theta$.

As explained above, the Fisher information for parameter $i$ is computed as follows:

$$F_i = \frac{1}{N} \sum_j^N \left( \frac{d}{d\theta_i} log\left[ p\left(X_j|\theta\right) \right] \right)^2 \tag{2}$$

Here:

- $N$ is the number of samples in the dataset A.
- $X_j$ represents the sample $j$ of dataset A.
- $p\left(X_j|\theta\right)$ represents log-likelihood, the probability of sample $j$ given the model parameters (optimized for task A).

The interest here is to know how much information does each parameter from the model trained on task A provide about the dataset A. Knowing the definition of the Fisher information, we can compute the parameter importance as shown in Equation 2. In practice, $log\left[ p\left(X_j|\theta\right) \right]$ is computed as the logarithm of the output probabilities of the model trained on task A. Then, the squared $1^{st}$ derivative of these log-probabilities are computed with respect to the optimal parameters for task A and averaged across all samples in the dataset. This way, an importance value for each parameter in the model is obtained to be used in the loss as part of the penalty term.

The advantage of this method is that the parameter importance for the source domain can be computed after training on the source domain with a forward pass of the whole dataset (without parameter update) and then, this source dataset is not needed anymore for training on the target domain. Additionally, the only extra memory required is to store the importance scores and optimal parameters from the source task.

#### *3.4.2. Memory Aware Synapses (MAS)*

This method, proposed by Aljundi et al. (2018), shares the core principle of EWC: constraining parameters crucial for the previous task (source domain) to stay close to their optimal values during training on

the new domain (target domain). However, MAS differs from EWC in how it calculates the parameter importance $\Omega$. Instead of relying on the Fisher information matrix, MAS computes the importance based on the average gradient of the network's output probability $l_2$ norm with respect to the optimal parameters for task A across all samples in the dataset A (see Equation 3). This essentially reflects how sensitive the network's output function is to changes in specific parameters. These importance values are incorporated into the loss function in the same way as in EWC, as shown in Equation 1.

$$\Omega_i = \frac{1}{N}\sum_{j}^{N}\left\|\frac{d}{d\theta_i} l_2\left(p\left(X_j|\theta\right)\right)\right\| \tag{3}$$

#### 3.4.3. Importance-based parameter-specific learning rate

Özgün et al. (2020) utilized the parameter importance to define a learning rate specific to each parameter, rather than adding a penalty term to the loss function. This means that parameters that are important for the previous task (higher importance) will have a reduced learning rate during training on the target domain. Conversely, less important parameters can adapt more freely to the new task. The parameter-specific learning rate for parameter $i$ in domain $d$ is defined as:

$$\alpha_i^d = (1 - \Omega_i)\alpha^d \tag{4}$$

Here, $\alpha^d$ represents the base learning rate for training on domain $d$, $\Omega_i$ represents the importance of parameter $i$ and $\alpha_i^d$ represents the specific learning rate for parameter $i$ and domain $d$. This approach offers computational efficiency as it avoids calculating an additional loss term and does not require to tune the hyperparameter $\lambda$ (present in EWC and MAS). In this work, this parameter-specific learning rate was evaluated employing the parameter importance computed both in EWC and MAS, and will be refered to as EWC-LR and MAS-LR, respectively.

#### 3.4.4. Experimental setup

To evaluate the effectiveness of CL techniques in mitigating catastrophic forgetting during domain adaptation, the investigation started by identifying the most suitable approach among the four presented methods: EWC, MAS, EWC-LR and MAS-LR. As it will be analyzed in the obtained results, EWC was the method which better preserved previous knowledge while allowing the network to adapt to the target domain. For this reason, it was selected for further study.

To better understand how the penalization weight affected the performance on both the source and target domains, different values of the hyperparameter $\lambda$ were tried (0.001, 0.01, 0.1, 1, 10, 10 and 1000). The optimal $\lambda$ value was selected based on a trade-off between learning the target domain and preserving the source domain knowledge, and was kept fixed for the rest of the experiments.

The following step was to assess the impact of the number of training images on the effectiveness of EWC. Similar to the one-shot and few-shot learning experiments in TL, fine-tuning with EWC was performed using 1, 2, 3, 5, 10, or all available images from the VH dataset. For consistency in result comparison, the patch sampling strategy and training hyperparameters employed during the EWC experiments were the same as those used in the TL experiments. This allowed for a direct comparison of the effectiveness of EWC against TL approaches.

### 3.5. Evaluation metrics

Perfect delineation of lesions is not always clinically relevant; however, accurate detection and localization of lesions are crucial for diagnosis and treatment planning. Given the clinical importance of lesion detection, lesion-wise metrics were prioritized over voxel-wise segmentation metrics, indicated by subindices $d$ and $s$, respectively. The employed evaluation metrics include:

- Dice Similarity Coefficient (DSC).

$$DSC = \frac{2 \cdot TP}{FN + FP + 2 \cdot TP} \tag{5}$$

  This metric was evaluated in two ways:

  - Voxel-wise DSC ($DSC_s$): measures the overlap between the predicted segmentation and the ground truth at the voxel level.
  - Lesion-wise DSC ($DSC_d$): assesses the overlap between the predicted lesions and the ground truth as whole objects.

- Detection True Positive Fraction ($TPF_d$): this metric counts the number of lesions correctly identified by the model. It was also computed per lesion size to analyze the model's performance across different lesion scales: small (1-10 voxels), medium (11-50 voxels) and large lesions (> 50 voxels).
- Detection False Positive Fraction ($FPF_d$): this metric counts the number of lesions incorrectly identified by the model.

$$TPF_d = \frac{TP}{TP + FN} \quad FPF_d = \frac{FP}{FP + TP} \tag{6}$$

- Detection F-score: this metric combines the precision and sensitivity for lesion detection, providing a more balanced evaluation between TP and FP.

$$F - score_d = \frac{TP}{TP + 0.5 \cdot (FP + FN)} \tag{7}$$

Table 3: Segmentation and detection results of the baseline models trained on each of the available datasets. The models trained on public datasets (Shifts and WMH2017) were also tested in the other datasets to assess the impact of domain shift.

| Training dataset | Inference dataset | $T_{bin}$ | $DSC_s$ | $DSC_d$ | $TPF_d$ | $FPF_d$ | $F-score_d$ |
|---|---|---|---|---|---|---|---|
| Shifts | Shifts | 0.4 | 0.635 ± 0.151 | 0.657 ± 0.139 | 0.665 ± 0.171 | 0.286 ± 0.197 | 0.657 ± 0.139 |
| | WMH2017 | 0.2 | 0.503 ± 0.210 | 0.545 ± 0.165 | 0.765 ± 0.102 | 0.544 ± 0.191 | 0.545 ± 0.165 |
| | VH | 0.4 | 0.359 ± 0.216 | 0.437 ± 0.192 | 0.601 ± 0.195 | 0.628 ± 0.204 | 0.437 ± 0.192 |
| WMH2017 | Shifts | 0.4 | 0.498 ± 0.149 | 0.591 ± 0.133 | 0.597 ± 0.184 | 0.319 ± 0.212 | 0.591 ± 0.133 |
| | WMH2017 | 0.2 | 0.743 ± 0.107 | 0.751 ± 0.101 | 0.810 ± 0.092 | 0.281 ± 0.117 | 0.752 ± 0.074 |
| | VH | 0.4 | 0.342 ± 0.202 | 0.472 ± 0.193 | 0.629 ± 0.196 | 0.594 ± 0.207 | 0.472± 0.193 |
| VH | | 0.4 | 0.506 ± 0.168 | 0.610 ± 0.167 | 0.599 ± 0.184 | 0.338 ± 0.193 | 0.610 ± 0.167 |

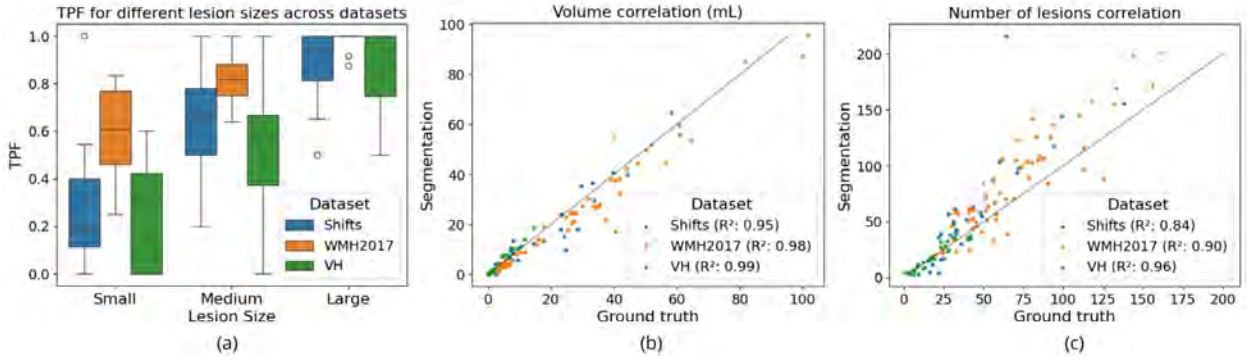



Figure 5: Results of the models trained and infered in the same dataset: $Shifts \rightarrow Shifts$, $WMH2017 \rightarrow WMH2017$, $VH \rightarrow VH$ (baseline). (a) TPF per lesion size across datasets. (b) Lesion volume (mL) correlation between segmentation and ground truth. (c) Number of lesions correlation between segmentation and ground truth.

### *3.6. Implementation details*

In this project, PyTorch version 2.2.0 (Paszke et al., 2019) and PyTorch Lightning version 2.2.1 (Falcon, William and The PyTorch Lightning team) were employed for deep learning model development and training. Code execution and hardware acceleration were achieved using CUDA version 12.1. The computational resources for this work were provided by a server equipped with three NVIDIA GeForce GTX 1080 Ti GPUs with 12Gb of memory each.

## 4. Results

### *4.1. Baseline: MS lesion segmentation*

The results obtained with the baseline models trained on each individual dataset (Shifts, WMH2017 and VH) can be seen in Table 3. The WMH2017 and Shifts models were tested in the same dataset they where trained on and in the other two datasets. Both models exhibited a significant drop in performance when evaluated on datasets different from their training data compared to the performance of the models trained with the same dataset. This confirmed the presence of a domain shift between the datasets.

Further analysis was conducted to understand the model's behavior across different lesion sizes. Boxplot Figure 5(a) depicted the sensitivity of each model for small, medium and large lesions. These results revealed variations in sensitivity based on lesion size. For all the models, the sensitivity increased with the lesion size. For small lesions, the mean sensitivity was lower compared to bigger lesions and with higher variability across the different images in the dataset.

The correlation between the models' performance and the ground truth is shown in Figures 5(b) and 5(c). Figure 5(b) presents a correlation plot between the volume of lesions detected by the models and the volume of lesions in the ground truth. This visualization helped assess the model's ability to accurately segment lesion sizes (segmentation). Similarly, Figure 5(c) shows the correlation between the number of lesions detected and the actual number of lesions present in the ground truth, providing insights into the model's lesion detection performance. Analyzing these correlations, it was observed that the WMH2017 model exhibited a tendency to undersegment lesions (Figure 5(b)), while detecting a higher number of false positives (Figure 5(c)). This suggested that the model might be capturing a higher number of small false positive lesions. In contrast, the analysis of the other two datasets (Shifts and VH) did not reveal any clear trends regarding undersegmentation or overdetection of lesions.

Qualitative segmentation examples from each model are showcased in Figure 6. For each example, the Figure displays the FLAIR image, the ground truth segmentation (overlaid in red), and the model's segmentation (overlaid in green). Even though T1w images were also employed for the segmentation, only FLAIR im-

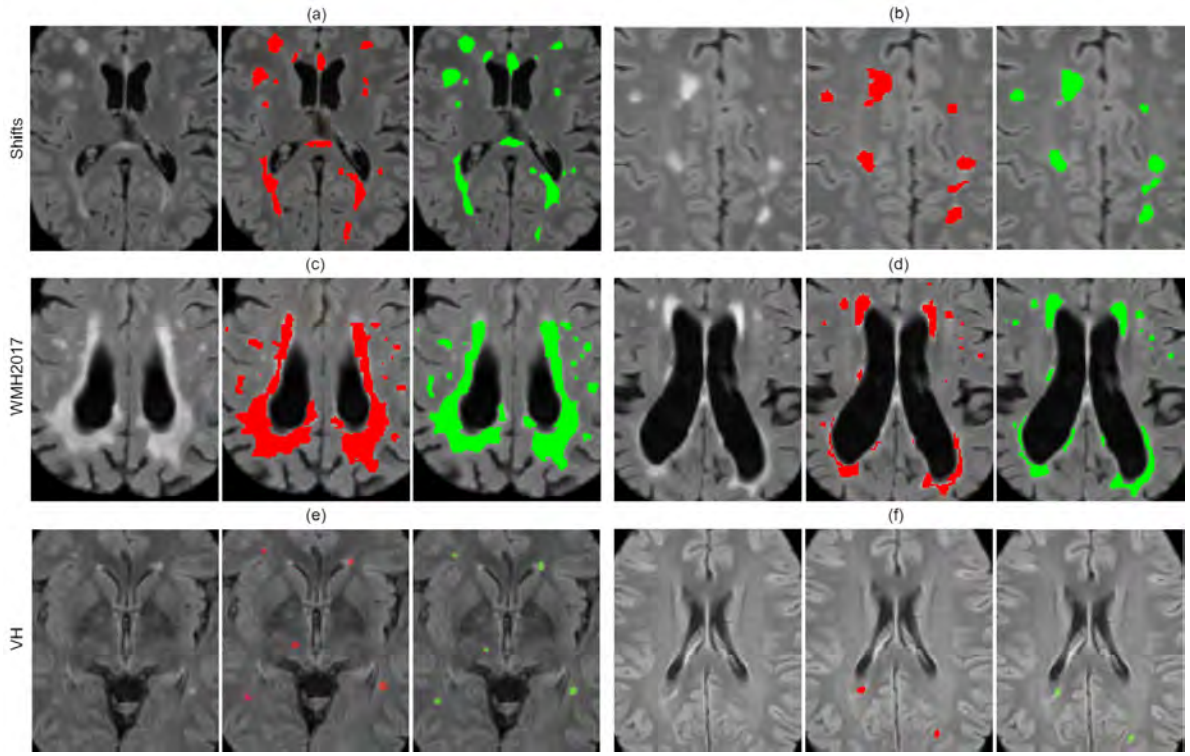



Figure 6: Qualitative results of the baseline models trained on three datasets: Shifts (a, b), WMH2017 (c, d) and VH (e,f). For each example, at the left is displayed the FLAIR scan, in the middle the ground truth is overlayed in red and in the right the model's segmentation is overlayed in green.

ages are included in the results Figures because of their good contrast between lesions and healthy tissue.

Based on the overall performance in Table 3, the WMH2017 model was chosen as the baseline for subsequent TL and CL experiments. This selection was justified by its good performance and the fact that the WMH2017 dataset was preprocessed in-house, allowing for greater control compared to the pre-processed Shifts dataset. Furthermore, Table 3 demonstrates a significant performance gap between the WMH2017 model's performance on the VH dataset and the upper bound (the model trained directly on VH data) (p-value $< 0.01$). More specifically, even though the sensitivity is even higher than the one of the model trained on the VH dataset, it results in a high number of false positive lesions, leading to a high gap in the detection F-score. This gap signifies the potential for improvement achievable through TL and CL strategies, that will be explored in the following sections.

### *4.2. Transfer learning for domain adaptation*

Table 4 presents the results of the analysis investigating which sections of the U-Net architecture were most effective for unfreezing during domain adaptation, as mentioned in Section 3.3. The Table 4 also includes results for the lower and upper bounds. The lower bound represents the performance of the WMH2017 model directly applied to the VH dataset without any adaptation (WMH2017 $\rightarrow$ VH), while the upper bound corresponds to the performance of a model trained directly on the VH dataset (VH $\rightarrow$ VH). Analysis of these results revealed that unfreezing only one layer, either from the encoder or the decoder, did not significantly improve the model's ability to adapt to the target domain (VH) (p-value $> 0.05$ in both cases). This limited adaptation is reflected in the persistence of a high number of false positive lesions detected by the model. Conversely, unfreezing all layers resulted in a significant improvement (p-value $< 0.01$), reaching a performance very similar to the model directly trained on the VH dataset (upper bound). Finally, unfreezing two layers, either from the encoder or the decoder led to a decrease in the false positive ratio compared to the baseline model (p-value $<$ 0.01 in both cases), although not as substantial as the complete unfreezing approach. Based on these findings, the subsequent few-shot domain adaptation experiments were conducted using models with the best unfreezing configurations: all layers unfrozen, or two layers unfrozen from either the encoder or the decoder.

#### *4.2.1. Few-shot transfer learning*

Figure 7 shows the influence of the target domain training set size and the unfreezing strategy on model performance. The experiment evaluated the WMH2017 model (with the best unfreezing configurations) retrained on the VH dataset using 1, 3, 5, 10, or all available images. The results for one-shot domain adaptation (1 training image) revealed that unfreezing only two lay-

Table 4: Results of the unfreezing tests during TL. The Table shows the results of fine tuning different layers of the WMH2017 model with all the training images of the VH dataset. The last rows represent the reference results: VH → VH is the upper bound (model trained and tested on VH) and WMH2017 → VH is the lower bound (model trained on WMH2017 and tested on VH).

| Unfreezed layers | $T_{bin}$ | $DSC_s$ | $DSC_d$ | $TPF_d$ | $FPF_d$ | $F-score_d$ |
|---|---|---|---|---|---|---|
| 1 layer encoder | 0.4 | 0.467 ± 0.193 | 0.486 ± 0.178 | 0.631 ± 0.145 | 0.575 ± 0.193 | 0.486 ± 0.178 |
| 2 layers encoder | 0.4 | 0.506 ± 0.185 | 0.569 ± 0.191 | 0.596 ± 0.206 | 0.443 ± 0.199 | 0.569 ± 0.191 |
| 1 layer decoder | 0.4 | 0.442 ± 0.193 | 0.501 ± 0.181 | 0.624 ± 0.191 | 0.555 ± 0.196 | 0.501 ± 0.181 |
| 2 layers decoder | 0.4 | **0.509 ± 0.183** | 0.608 ± 0.174 | **0.643 ± 0.191** | 0.392 ± 0.208 | 0.608 ± 0.174 |
| All layers | 0.4 | 0.507 ± 0.178 | **0.634 ± 0.152** | 0.635 ± 0.181 | **0.347 ± 0.168** | **0.634 ± 0.152** |
| VH → VH | 0.4 | 0.506 ± 0.168 | 0.610 ± 0.167 | 0.599 ± 0.184 | 0.338 ± 0.193 | 0.610 ± 0.167 |
| WMH2017 → VH | 0.4 | 0.342 ± 0.202 | 0.472 ± 0.193 | 0.629 ± 0.196 | 0.594 ± 0.207 | 0.472 ± 0.193 |

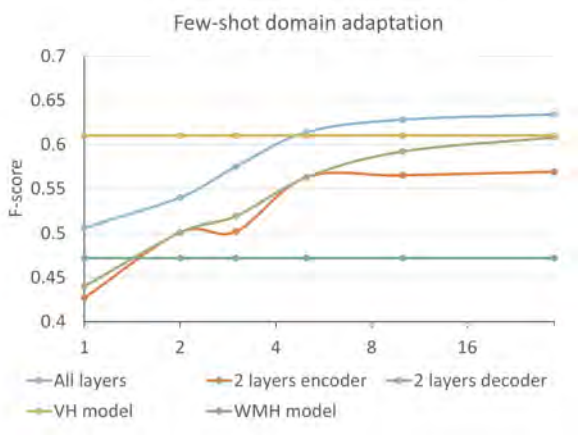



Figure 7: Few-shot domain adaptation results on the VH dataset (target domain). Detection F-score for different number of training images of the VH dataset (1, 2, 3, 5, 10 or all the available) unfreezing different sections of the network trained on WMH2017 (2 layers of the encoder or decoder, or all the layers). The lower bound corresponds to the model trained on WMH2017 and tested on VH (WMH model) and the upper bound corresponds to the model trained and tested on VH (VH model).

ers of the encoder or decoder led to a decrease in performance compared to the lower bound (WMH2017 model directly applied to VH). In contrast, unfreezing all layers in the one-shot scenario already improved the results compared to the base model and achieved similar performance to training with 3 images when using the other unfreezing strategies (two layers unfrozen in encoder or decoder). When unfreezing all the layers, training with only 5 images was enough to reach a performance comparable to the upper bound (model trained directly on VH). Given that retraining all the layers yielded the best results in both the few-shot experiments and full training scenario, this strategy was chosen for the subsequent CL analysis.

### *4.3. Continuous learning*

While TL offered a solution to domain shift by leveraging a pre-trained model on a source domain (WMH2017) for adaptation to a new target domain (VH), it suffered from catastrophic forgetting, leading to a significant performance drop in the source domain (p-value < 0.01).

The next step focused on mitigating catastrophic forgetting during domain adaptation. To achieve this, the first step involved identifying the most suitable method among various CL strategies based on prior-focused regularization. Four methods were quantitatively evaluated: EWC, MAS, EWC-LR and MAS-LR. Table 5 summarizes the results obtained when retraining the entire WMH2017 model using all VH images. The Table presents performance metrics for both the source and target domains, along with the optimal $\lambda$ values for MAS and EWC (the selection of the optimal $\lambda$ was explained in Section 3.4.4). Notably, EWC-LR and MAS-LR did not require hyperparameter tuning for this specific learning rate adaptation strategy. These results revealed that EWC emerged as the most effective method in preserving the source domain knowledge, with a significant improvement with respect to TL techniques (p-value < 0.01), so it was chosen as the preferred strategy for further CL experiments.

To understand how the penalization weight ($\lambda$) in EWC affected performance on the source and target domains, different $\lambda$ values were explored. Figure 8 depicts the results of these experiments. The graph reveals a trade-off between source domain knowledge preservation and target domain learning flexibility. This is reflected in the tendency for higher values of the hyperparameter $\lambda$ to correspond with greater preservation of source domain knowledge, but also with a potentially more restricted ability for the model to learn and adapt to the target domain.

Figure 9 provides qualitative results to visualize the performance differences between the baseline model, TL strategies, and CL with EWC in both the source and target domains. The upper section of the Figure shows the source domain results. Here, the baseline model achieved good performance with a segmentation very close to the ground truth. However, due to catastrophic forgetting, the TL strategies exhibited a high number of false positive lesions in the source domain. In the target domain, the baseline model showed poor performance due to the domain shift, again evident by the presence of numerous false positive lesions. Following a TL strategy, the model successfully adapted to the target domain, achieving a segmentation comparable to

Table 5: Comparison of different prior-focused regularization-based CL methods. The upper row shows the results of TL as a reference.

| | | Target domain (VH) | | | | | Source domain (WMH2017) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CL method | Lambda | $T_{bin}$ | $DSC_d$ | $TPF_d$ | $FPF_d$ | $F-score_d$ | $T_{bin}$ | $DSC_d$ | $TPF_d$ | $FPF_d$ | $F-score_d$ |
| Naïve TL | - | 0.4 | 0.634 | 0.635 | 0.347 | 0.634 | 0.1 | 0.544 | 0.583 | 0.465 | 0.544 |
| **EWC** | **0.1** | **0.4** | **0.620** | **0.599** | **0.301** | **0.620** | **0.1** | **0.627** | **0.610** | **0.342** | **0.630** |
| MAS | 1 | 0.4 | 0.608 | 0.639 | 0.400 | 0.608 | 0.1 | 0.614 | 0.715 | 0.450 | 0.614 |
| EWC - LR | - | 0.4 | 0.648 | 0.579 | 0.237 | 0.648 | 0.1 | 0.609 | 0.564 | 0.318 | 0.609 |
| MAS - LR | - | 0.4 | 0.630 | 0.556 | 0.238 | 0.630 | 0.1 | 0.599 | 0.571 | 0.356 | 0.599 |

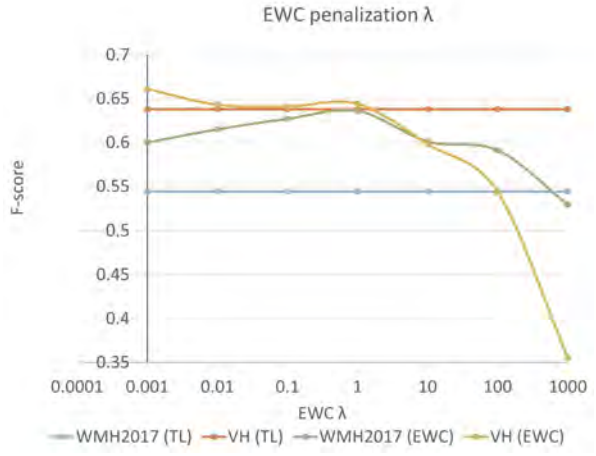



Figure 8: Detection F-score for different values of $\lambda$ in EWC penalization, for both the target domain (VH, in yellow) and the source domain (WMH2017, in green). Also, the results of TL are included for reference, for both the target domain (orange) and the source domain (blue).

the ground truth. Finally, EWC results demonstrate the ability of this technique to balance the trade-off between source and target domain performance. Compared to TL, EWC preserves source domain knowledge, resulting in segmentation more similar to the baseline model, while still adapting to the new target domain.

#### *4.3.1. Few-shot continuous learning*

Having established EWC's effectiveness in preserving source domain knowledge, the same few-shot TL experiments were replicated, but with EWC penalty applied ($\lambda$ = 0.1 was selected as the optimal value for all the experiments). Figure 10 presents the results for both the source and target domains. The results demonstrated that EWC did not prevent the network from learning the target domain, as the performance was very similar to that achieved with regular TL. Importantly, this method successfully alleviated catastrophic forgetting, reducing the performance drop in the source domain compared to TL alone. Additionally, EWC provided more stable performance in the source domain when compared to TL, where the extent of forgetting seemed unpredictable.

## 5. Discussion

This work investigated different techniques to solve the domain shift problem in the context of MS lesion segmentation, while mitigating catastrophic forgetting in the source domain. The results demonstrated the effectiveness of these approaches in improving model performance on a target domain (VH) while preserving knowledge from a source domain (WMH2017).

The baseline models trained on each public dataset (WMH2017 and Shifts) achieved detection results comparable to the state-of-the-art. For instance, on the WMH2017 dataset, the baseline model achieved a TPF of 0.810 and an F-score of 0.752. These values are competitive with the winning entry in the corresponding segmentation challenge, which reported a TPF of 0.84 and an F-score of 0.76 (Kuijf et al., 2019). Moreover, these baseline models confirmed the presence of a domain shift between the datasets. This was evident from the drop in performance observed when evaluating the models on datasets different from their training data (Table 3). The analysis of lesion size sensitivity (Figure 5(a)) highlighted the challenges associated with detecting small lesions (1-10 voxels), where all models exhibited lower sensitivity compared to larger lesions.

The TL experiments explored unfreezing different sections of the U-Net architecture during fine-tuning on the VH dataset (Table 4). Unfreezing one layer, either from the encoder or the decoder, did not significantly improve adaptation, likely because the model lacked enough flexibility to learn the new domain characteristics. While unfreezing two layers (from the encoder or decoder) resulted in a better adaptation to the source domain, the best configuration was to unfreeze all layers, which resulted in performance comparable to training directly on VH data. This suggests that for larger domain shifts, extensive adaptation of the model is necessary.

The investigation into the impact of the target training set size (Figure 7) revealed that unfreezing all layers yielded the best results in both few-shot and full training scenarios. In this case, employing only 5 training images was enough to adapt to the target domain. In contrast, one-shot domain adaptation with partial unfreezing led to a decrease in performance compared to the lower bound, suggesting that finetuning such a small

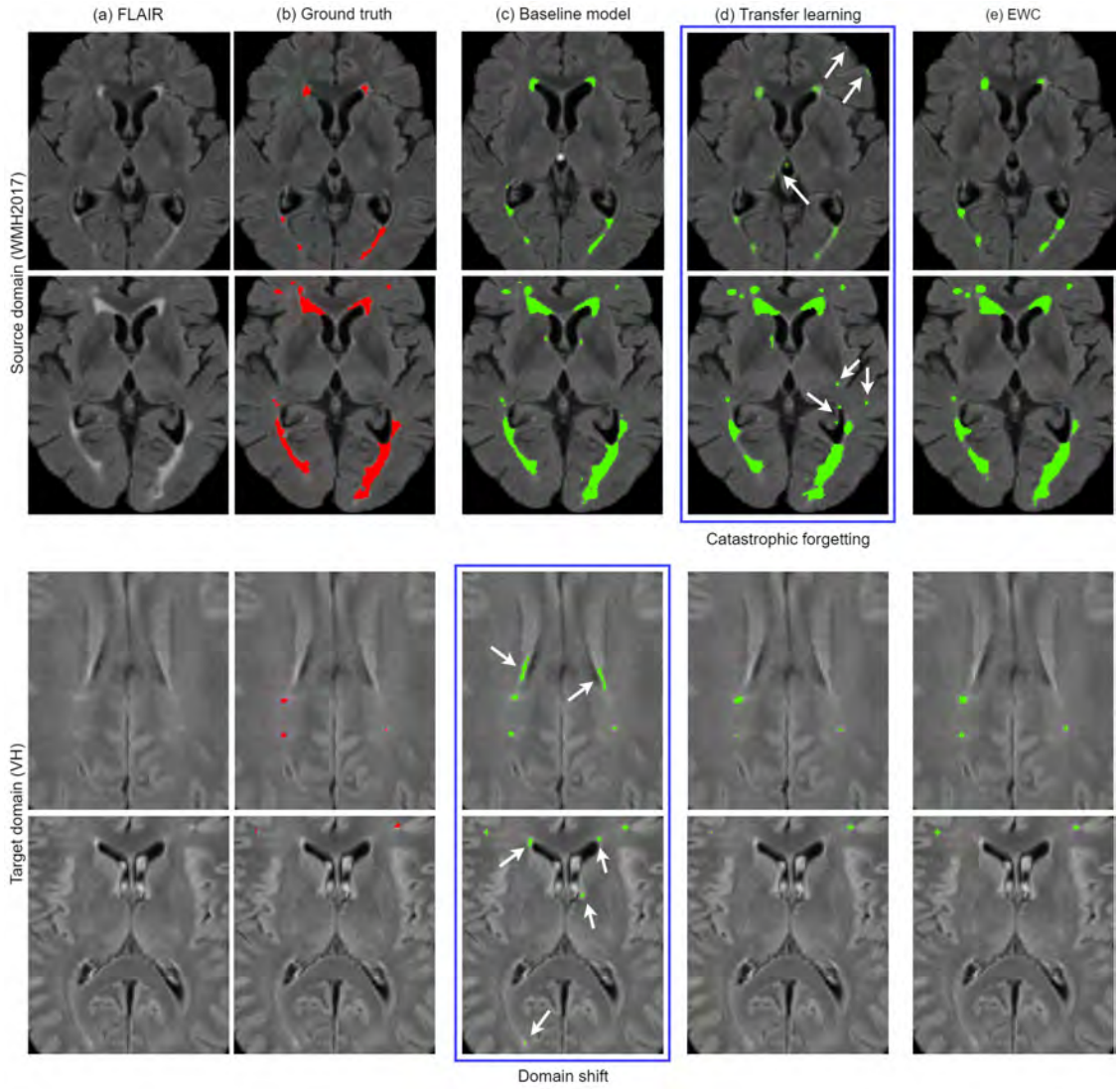



Figure 9: Comparison of TL and CL segmentation results on the source and target domain. (a) FLAIR image. (b) Ground truth. (c) **Baseline model**: effective on the source domain but it fails to generalize to the target domain (domain shift). (d) **TL** achieves successful adaptation to the target domain, but suffers from catastrophic forgetting on the source domain. (e) **EWC** shows good performance in both the source and target domains.

number of parameters might hinder effective adaptation with such a small training set. For small number of training images, the results of unfreezing two layers of the encoder or the decoder were similar. However, with the full training set, opposed to what Shirokikh et al. (2020) stated, finetuning the decoder resulted in better adaptation to the VH dataset compared to finetuning the encoder. This might be due to the fact that preserving the pre-trained feature hierarchy of the encoder and only adjusting the reconstruction for the new domain allowed the model to focus on the specific task requirements of the VH dataset without disrupting the general feature extraction capabilities learned during pre-training. This can lead to faster adaptation and potentially better performance on the new domain.

Among the evaluated CL techniques (EWC, MAS, EWC-LR, and MAS-LR), EWC emerged as the most effective method for preserving knowledge from the source domain (WMH2017) while still adapting to the target domain (VH) (Table 5). The main advantage of EWC-LR and MAS-LR is that these methods do not require tuning the hyperparameter $\lambda$. However, the fact that the parameter-specific learning rate is computed before training based on the parameter's importance for the source domain, makes it less adaptive than MAS and EWC. These methods adapt the penalization during training at each step, based on both the importance of the parameter and the difference between its current value (training on the target domain) and optimal value (trained on the source domain), leading to a more flexible technique. MAS has shown promising results but, in this case EWC led to slightly better performance on

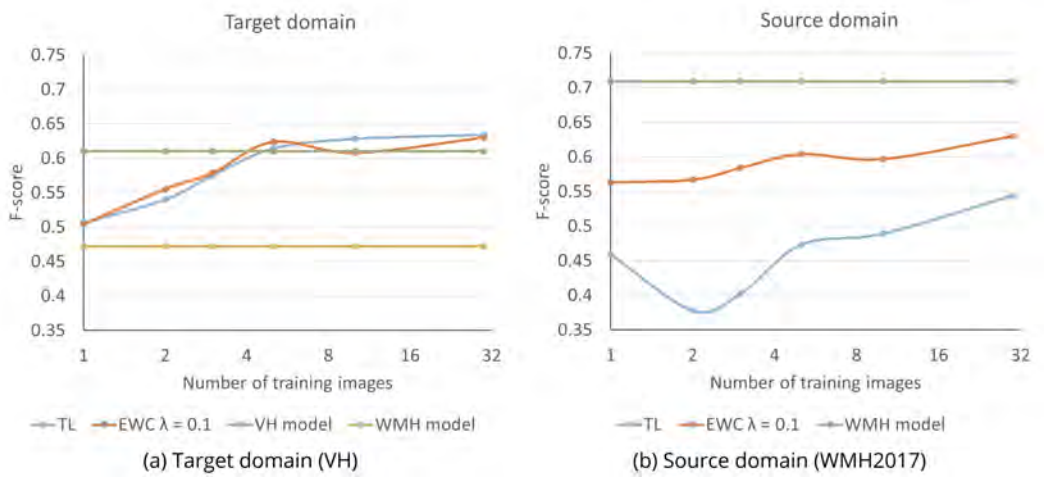



Figure 10: Few-shot domain adaptation comparison between TL and EWC with $\lambda = 0.1$. (a) **Target domain** (VH): the upper bound is the model trained and tested on VH (VH model), and the lower bound corresponds to the model trained on WMH2017 and tested on VH (WMH model). (b) **Source domain** (WMH2017): the upper bound corresponds to the model trained and tested on WMH2017.

both the source and target domain, and the best performance among the four methods on the source domain. The exploration of the penalization weight ($\lambda$) in EWC (Figure 8) emphasized the importance of careful hyperparameter tuning for balancing source domain knowledge preservation and target domain learning flexibility. As the $\lambda$ value increased, the model allowed less flexibility for learning the target domain. Conversely, for the source domain, a higher $\lambda$ value translated to better knowledge preservation. However, excessively high $\lambda$ led to a decrease in source domain performance. This was likely because a very strong penalty resulted in high loss values during training, which in turn led to significant changes in model parameters, hindering the desired behavior of preserving previous knowledge. Careful tuning of this parameter is therefore crucial.

The few-shot TL experiments with EWC (Figure 10) demonstrated that EWC successfully mitigated catastrophic forgetting compared to regular TL. This was evident from the reduced performance drop in the source domain (WMH2017) when using EWC. Additionally, EWC provided more stable performance in the source domain across different training set sizes compared to TL alone. The qualitative results (Figure 9) further supported these findings.

### *5.1. Limitations and future work*

While this work demonstrated the potential of EWC for mitigating catastrophic forgetting in domain adaptation scenarios for MS lesion segmentation, some limitations need to be addressed in future studies.

One key limitation is the careful tuning required for the EWC penalty term ($\lambda$). While effective in this instance, the optimal value for $\lambda$ may not generalize well to other datasets, particularly those with varying levels of domain shift or lesion characteristics. This highlights the need to explore more robust hyperparameter selection techniques or even investigate alternative CL methods that are less reliant on manual tuning.

Even though the proposed work achieved a good balance between source domain knowledge preservation and target domain adaptation, there is still room for improvement. One potential future work is to explore domain adaptation techniques specifically designed to reduce the initial domain shift between source and target data. Additionally, focusing on advanced preprocessing techniques that normalize image intensities or address potential artifacts across datasets could further enhance model performance.

Finally, the ability to achieve good results in both the source and target domain with few images is promising. However, reducing the amount of necessary images even further would enhance the practicality of these methods in real-world settings where acquiring large amounts of labeled target data might be challenging. Here, investigating modifications to patch sampling techniques during training could be beneficial. Additionally, a more extensive study is necessary to solidify these findings. Repeating the few-shot learning experiments with various image combinations would provide valuable insights into how sensitive the results are to the specific selection of images in such scenarios. This would provide a more robust understanding of the generalizability of the approach with minimal source domain data.

By addressing these limitations and exploring the proposed future directions, the effectiveness of this approach for CL in MS lesion segmentation tasks with domain shift can be further strengthen.

## 6. Conclusions

In this work, a deep learning framework based on a U-Net architecture was developed for MS lesions segmentation. Several public international and in-house datasets were used to evaluate the performance of the model on in- and out-domain images, to assess the domain shift problem. TL techniques were studied for domain adaptation, and CL techniques were explored to mitigate catastrophic forgetting suffered by TL methods. For both CL and TL techniques, few-shot approaches were analyzed due to the high interest in reducing the number of training images needed for domain adaptation.

The analysis of the baseline model revealed a significant drop in performance when testing the source model on the target domain, with an average F-score decrease of around 14%. While TL achieved full-domain adaptation with only 5 target images, using only 3 images led to an F-score improvement of almost 10%.

Furthermore, EWC demonstrated its effectiveness in mitigating catastrophic forgetting. Notably, this work was the first to apply EWC for domain-incremental learning in MS lesion segmentation. Results showed that the source model's performance drop during adaptation to the target domain using TL ranged from 20% to 37% in terms of F-score. EWC successfully reduced catastrophic forgetting by 8% to 19% across different training set sizes. Moreover, EWC achieved comparable performance on the target domain as TL techniques, even in few-shot and full-training settings.

Finally, a significant advantage of EWC is its efficiency. It enables adaptation without requiring source domain images during target domain training. This not only translates to memory efficiency but also addresses critical privacy concerns within the medical domain. Additionally, EWC avoids introducing new network parameters, eliminates the need for domain labels during inference, and does not require training separate neural networks. These characteristics make EWC a promising approach for real-world applications, particularly for companies that operate in the medical image analysis sector.

## Acknowledgments

I would like to thank my supervisors Dr Xavier Lladó and Dr Sergi Valverde for their guidance, support and expertise during the development of this Masther Thesis and for giving me the opportunity to join the ViCOROB institute and Tensor Medical company. Furthermore, I would like to express my appreciation to the entire Tensor Medical team for their warm welcome and continuous support. Finally, thanks to my family and friends for accompanying me through this experience.

*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

Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., Tuytelaars, T., 2018. Memory aware synapses: Learning what (not) to forget, in: Computer Vision – ECCV 2018, Springer International Publishing. pp. 144–161.

Ashtari, P., Barile, B., Van Huffel, S., Sappey-Marinier, D., 2022. New multiple sclerosis lesion segmentation and detection using pre-activation u-net. Frontiers in Neuroscience 16, 975862. doi:10.3389/fnins.2022.975862.

Aslani, S., Dayan, M., Storelli, L., Filippi, M., Murino, V., Rocca, M.A., Sona, D., 2019. Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. NeuroImage 196, 1–15. doi:10.1016/j.neuroimage.2019.03.068.

Baweja, C., Glocker, B., Kamnitsas, K., 2018. Towards continual learning in medical imaging. arXiv: 1811.02496 .

Bayasi, N., Hamarneh, G., Garbi, R., 2021. Culprit-prune-net: Efficient continual sequential multi-domain learning with application to skin lesion classification, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing. pp. 165–175. doi:10.1007/978-3-030-87234-2_16.

Chen, S., Tang, F., 2022. Breast cancer detection model training strategy based on continual learning, in: CAIBDA 2022; 2nd International Conference on Artificial Intelligence, Big Data and Algorithms, pp. 1–5.

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, Springer International Publishing. pp. 424–432. doi:10.1007/978-3-319-46723-8_49.

Commowick, O., Cervenansky, F., Cotton, F., Dojat, M., 2021. Msseg-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure, in: MICCAI 2021 - 24th International Conference on Medical Image Computing and Computer Assisted Intervention, Strasbourg, France.

Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., Pop, S.C., Girard, P., Améli, R., Ferré, J.C., Kerbrat, A., Tourdias, T., Cervenansky, F., Glatard, T., Beaumont, J., Doyle, S., Forbes, F., Knight, J., Khademi, A., Mahbod, A., Wang, C., McKinley, R., Wagner, F., Muschelli, J., Sweeney, E., Roura, E., Lladó, X., Santos, M.M., Santos, W.P., Silva-Filho, A.G., Tomas-Fernandez, X., Urien, H., Bloch, I., Valverde, S., Cabezas, M., Vera-Olmos, F.J., Malpica, N., Guttmann, C., Vukusic, S., Edan, G., Dojat, M., Styner, M., Warfield, S.K., Cotton, F., Barillot, C., 2018. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. Scientific Reports 8, 13650. doi:10.1038/s41598-018-31911-7.

Compston, A., Coles, A., 2008. Multiple sclerosis. The Lancet 372, 1502–1517. doi:10.1016/S0140-6736(08)61620-7.

Coupe, P., Yger, P., Prima, S., Hellier, P., Kervrann, C., Barillot, C., 2008. An optimized blockwise nonlocal means denoising filter for 3-d magnetic resonance images. IEEE Transactions on Medical Imaging 27, 425–441. doi:10.1109/TMI.2007.906087.

Falcon, William and The PyTorch Lightning team, . Pytorch lightning (version 2.2.1). URL: `https://www.pytorchlightning.ai`.

Feng, Y., Pan, H., Meyer, C., Feng, X., 2019. A self-adaptive network for multiple sclerosis lesion segmentation from multi-contrast mri with various imaging sequences, in: 2019 IEEE 16th International

Symposium on Biomedical Imaging (ISBI 2019), pp. 472–475. doi:10.1109/ISBI.2019.8759522.

Fenneteau, A., Bourdon, P., Helbert, D., Fernandez-Maloigne, C., Habas, C.N., Guillevin, R., 2021. Investigating efficient CNN architecture for multiple sclerosis lesion segmentation. Journal of Medical Imaging 8. doi:10.1117/1.JMI.8.1.014504.

Filippi, M., Preziosa, P., Banwell, B.L., Barkhof, F., Ciccarelli, O., Stefano, N.D., Geurts, J.J.G., Paul, F., Reich, D.S., Toosy, A.T., Traboulsee, A., Wattjes, M.P., Yousry, T.A., Gass, A., Lubetzki, C., Weinshenker, B.G., Rocca, M.A., 2019. Assessment of lesions on magnetic resonance imaging in multiple sclerosis: practical guidelines. Brain 142, 1858–1875. doi:10.1093/brain/awz144.

Filippi, M., Preziosa, P., Meani, A., Dalla Costa, G., Mesaros, S., Drulovic, J., Ivanovic, J., Rovira, A., Tintorè, M., Montalban, X., Ciccarelli, O., Brownlee, W., Miszkiel, K., Enzinger, C., Khalil, M., Barkhof, F., Strijbis, E.M., Frederiksen, J.L., Cramer, S.P., Fainardi, E., Amato, M.P., Gasperini, C., Ruggieri, S., Martinelli, V., Comi, G., Rocca, M.A., on behalf of the MAGNIMS Study Group, Stefano, N.D., Palace, J., Kappos, L., Sastre-Garriga, J., Yousry, T., 2022. Performance of the 2017 and 2010 revised mcdonald criteria in predicting ms diagnosis after a clinically isolated syndrome: A magnims study. Neurology 98. doi:10.1212/WNL.0000000000013016.

Functional Imaging Laboratory, U.C.L., 2014. Spm12. URL: `https://www.fil.ion.ucl.ac.uk/spm/software/spm12/`.

Gamal, R., Barka, H., Hadhoud, M., 2023. Gau u-net for multiple sclerosis segmentation. Alexandria Engineering Journal 73, 625–634. doi:10.1016/j.aej.2023.04.069.

van Garderen, K., van der Voort, S., Incekara, F., Smits, M., Klein, S., 2019. Towards continuous learning for glioma segmentation with elastic weight consolidation. arXiv:1909.11479 .

Ghafoorian, M., Mehrtash, A., Kapur, T., Karssemeijer, N., Marchiori, E., Pesteie, M., Guttmann, C.R.G., de Leeuw, F.E., Tempany, C.M., van Ginneken, B., Fedorov, A., Abolmaesumi, P., Platel, B., Wells, W.M., 2017. Transfer Learning for Domain Adaptation in MRI: Application in Brain Lesion Segmentation. Springer International Publishing. p. 516–524. doi:10.1007/978-3-319-66179-7_59.

Grahl, S., Pongratz, V., Schmidt, P., Engl, C., Bussas, M., Radetz, A., Gonzalez-Escamilla, G., Groppa, S., Zipp, F., Lukas, C., Kirschke, J., Zimmer, C., Hoshi, M., Berthele, A., Hemmer, B., Mühlau, M., 2019. Evidence for a white matter lesion size threshold to support the diagnosis of relapsing remitting multiple sclerosis. Multiple Sclerosis and Related Disorders 29, 124–129. doi:10.1016/j.msard.2019.01.042.

Guan, H., Liu, M., 2022. Domain adaptation for medical image analysis: A survey. IEEE Transactions on Biomedical Engineering 69, 1173–1185. doi:10.1109/TBME.2021.3117407.

Hashemi, M., Akhbari, M., Jutten, C., 2022. Delve into multiple sclerosis (ms) lesion exploration: A modified attention u-net for ms lesion segmentation in brain mri. Computers in Biology and Medicine 145, 105402. doi:10.1016/j.compbiomed.2022.105402.

hu, C., Kang, G., Hou, B., Ma, Y., Labeau, F., Su, Z., 2020. Acu-net: A 3d attention context u-net for multiple sclerosis lesion segmentation, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1384–1388. doi:10.1109/ICASSP40776.2020.9054616.

Isensee, F., Jaeger, P.F., Kohl, S.A.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. 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., Bendszus, M., Maier-Hein, K.H., Kickingereder, P., 2019. Automated brain extraction of multisequence mri using artificial neural networks. Human Brain Mapping 40, 4952–4964. doi:10.1002/hbm.24750.

Karani, N., Chaitanya, K., Baumgartner, C., Konukoglu, E., 2018. A lifelong learning approach to brain mr segmentation across scanners and protocols, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Springer International Publishing. pp. 476–484. doi:10.1007/978-3-030-00928-1_54.

Karimi, D., Warfield, S.K., Gholipour, A., 2021. Transfer learning in medical image segmentation: New insights from analysis of the dynamics of model parameters and learned representations. Artificial Intelligence in Medicine 116, 102078. doi:10.1016/j.artmed.2021.102078.

Karthik, E.N., Kerbrat, A., Labauge, P., Granberg, T., Talbott, J., Reich, D.S., Filippi, M., Bakshi, R., Callot, V., Chandar, S., Cohen-Adad, J., 2022. Segmentation of multiple sclerosis lesions across hospitals: Learn continually or train from scratch? arXiv: 2210.15091 .

Kaustaban, V., Ba, Q., Bhattacharya, I., Sobh, N., Mukherjee, S., Martin, J., Miri, M.S., Guetter, C., Chaturvedi, A., 2022. Characterizing continual learning scenarios for tumor classification in histopathology images, in: Medical Optical Imaging and Virtual Microscopy Image Analysis, Springer Nature Switzerland. doi:10.1007/978-3-031-16961-8_18.

Kirkpatrick, James, Pascanu, Razvan, Rabinowitz, Neil, Veness, Joel, Desjardins, Guillaume, Rusu, A., A., Milan, Kieran, Quan, John, Ramalho, Tiago, Grabska-Barwinska, Agnieszka, Hassabis, Demis, Clopath, Claudia, Kumaran, Dharshan, Hadsell, Raia, 2017. Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences 114, 3521–3526. doi:10.1073/pnas.1611835114.

Krishnamoorthy, S., Zhang, Y., Kadry, S., Yu, W., 2022. Framework to segment and evaluate multiple sclerosis lesion in mri slices using vgg-unet. Computational Intelligence and Neuroscience 2022, 1–10. doi:10.1155/2022/4928096.

Kuijf, H.J., Biesbroek, J.M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M.J., Casamitjana, A., Collins, D.L., Dadar, M., Georgiou, A., Ghafoorian, M., Jin, D., Khademi, A., Knight, J., Li, H., Lladó, X., Luna, M., Mahmood, Q., McKinley, R., Mehrtash, A., Ourselin, S., Park, B.Y., Park, H., Park, S.H., Pezold, S., Puybareau, E., Rittner, L., Sudre, C.H., Valverde, S., Vilaplana, V., Wiest, R., Xu, Y., Xu, Z., Zeng, G., Zhang, J., Zheng, G., Chen, C., van der Flier, W., Barkhof, F., Viergever, M.A., Biessels, G.J., 2019. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. IEEE Transactions on Medical Imaging 38, 2556–2568. doi:10.1109/TMI.2019.2905770.

Kumari, P., Chauhan, J., Bozorgpour, A., Huang, B., Azad, R., Merhof, D., 2024. Continual learning in medical image analysis: A comprehensive review of recent advancements and future prospects. arXiv: 2312.17004 .

Kushibar, K., Salem, M., Valverde, S., Rovira, A., Salvi, J., Oliver, A., Lladó, X., 2021. Transductive transfer learning for domain adaptation in brain magnetic resonance image segmentation. Frontiers in Neuroscience 15. doi:10.3389/fnins.2021.608808.

Li, K., Yu, L., Heng, P.A., 2023. Domain-incremental cardiac image segmentation with style-oriented replay and domain-sensitive feature whitening. IEEE Transactions on Medical Imaging 42, 570–581. doi:10.1109/TMI.2022.3211195.

Li, Z., Hoiem, D., 2018. Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence 40, 2935–2947. doi:10.1109/TPAMI.2017.2773081.

Liu, P., Wang, X., Fan, M., Pan, H., Yin, M., Zhu, X., Du, D., Zhao, X., Xiao, L., Ding, L., Wu, X., Zhou, S.K., 2022. Learning incrementally to segment multiple organs in a ct image, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, Springer Nature Switzerland. pp. 714–724. doi:10.1007/978-3-031-16440-8_68.

Lladó, X., Oliver, A., Cabezas, M., Freixenet, J., Vilanova, J.C., Quiles, A., Valls, L., Ramió-Torrentà, L., Àlex Rovira, 2012. Segmentation of multiple sclerosis lesions in brain mri: A review of automated approaches. Information Sciences 186, 164–185. doi:10.1016/j.ins.2011.10.011.

Malinin, A., Athanasopoulos, A., Barakovic, M., Cuadra, M.B., Gales, M.J.F., Granziera, C., Graziani, M., Kartashev, N., Kyriakopoulos, K., Lu, P.J., Molchanova, N., Nikitakis, A., Raina, V., Rosa, F.L., Sivena, E., Tsarsitalidis, V., Tsompopoulou, E., Volf,

E., 2022. Shifts 2.0: Extending the dataset of real distributional shifts. arXiv: 2206.15407 .

McClure, P., Kaczmarzyk, J.R., Ghosh, S.S., Bandettini, P., Zheng, C.Y., Lee, J.A., Nielson, D., Pereira, F., 2018. Distributed weight consolidation: A brain segmentation case study. Advances in neural information processing systems 31, 4093–4103.

Multiple Sclerosis International Federation, 2021. The multiple sclerosis international federation – atlas of ms – 3rd edition, part 2: clinical management of multiple sclerosis around the world. Open-source publication.

Ozdemir, F., Fuernstahl, P., Goksel, O., 2018. Learn the new, keep the old: Extending pretrained models with new anatomy and images, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Springer International Publishing. pp. 361–369. doi:10.1007/978-3-030-00937-3_42.

Özgün, S., Rickmann, A.M., Roy, A.G., Wachinger, C., 2020. Importance driven continual learning for segmentation across domains, in: Machine Learning in Medical Imaging, Springer International Publishing. pp. 423–433.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32. Curran Associates, Inc., pp. 8024–8035.

Perkonigg, M., Hofmanninger, J., Herold, C.J., Brink, J.A., Pianykh, O., Prosch, H., Langs, G., 2021. Dynamic memory to alleviate catastrophic forgetting in continual learning with medical imaging. Nature Communications 12, 5678. doi:10.1038/s41467-021-25858-z.

Petronella, Vincken, K.L., van Bochove, G.S., van Osch, M.J., van der Grond, J., 2005. Probabilistic segmentation of brain tissue in mr imaging. NeuroImage 27, 795–804. doi:10.1016/j.neuroimage.2005.05.046.

Pianykh, O.S., Langs, G., Dewey, M., Enzmann, D.R., Herold, C.J., Schoenberg, S.O., Brink, J.A., 2020. Continuous learning ai in radiology: Implementation principles and early applications. Radiology 297, 6–14. doi:10.1148/radiol.2020200038.

Ravishankar, H., Venkataramani, R., Anamandra, S., Sudhakar, P., Annangi, P., 2019. Feature transformers: Privacy preserving lifelong learners for medical imaging, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Springer International Publishing. pp. 347–355. doi:10.1007/978-3-030-32251-9_38.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. Springer International Publishing. volume 9351. p. 234–241. doi:10.1007/978-3-319-24574-4_28.

Roura, E., Oliver, A., Cabezas, M., Valverde, S., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Rovira, A., Lladó, X., 2015. A toolbox for multiple sclerosis lesion segmentation. Neuroradiology 57, 1031–1043. doi:10.1007/s00234-015-1552-2.

Salem, M., Ryan, M.A., Oliver, A., Hussain, K.F., Lladó, X., 2022. Improving the detection of new lesions in multiple sclerosis with a cascaded 3d fully convolutional neural network approach. Frontiers in Neuroscience 16, 1007619. doi:10.3389/fnins.2022.1007619.

Shirokikh, B., Zakazov, I., Chernyavskiy, A., Fedulova, I., Belyaev, M., 2020. First u-net layers contain more domain specific information than the last ones. arXiv: 2008.07357 .

Srivastava, S., Yaqub, M., Nandakumar, K., Ge, Z., Mahapatra, D., 2021. Continual domain incremental learning for chest x-ray classification in low-resource clinical settings, in: Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health, Springer International Publishing. pp. 226–238. doi:10.1007/978-3-030-87722-4_21.

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. doi:10.1109/TMI.2010.2046908.

Valverde, S., Cabezas, M., Roura, E., González-Villà, S., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Rovira, A., Oliver, A., Lladó, X., 2017. Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach. NeuroImage 155, 159–168. doi:10.1016/j.neuroimage.2017.04.034.

Valverde, S., Salem, M., Cabezas, M., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Àlex Rovira, Salvi, J., Oliver, A., Lladó, X., 2019. One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. NeuroImage: Clinical 21, 101638. doi:10.1016/j.nicl.2018.101638.

Van Leemput, K., Maes, F., Vandermeulen, D., Colchester, A., Suetens, P., 2001. Automated segmentation of multiple sclerosis lesions by model outlier detection. IEEE Transactions on Medical Imaging 20, 677–688. doi:10.1109/42.938237.

Yan, S., Xie, J., He, X., 2021. Der: Dynamically expandable representation for class incremental learning, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA. pp. 3013–3022. doi:10.1109/CVPR46437.2021.00303.

Zeng, C., Gu, L., Liu, Z., Zhao, S., 2020. Review of deep learning approaches for the segmentation of multiple sclerosis lesions on brain mri. Frontiers in Neuroinformatics 14. doi:10.3389/fninf.2020.610967.

Zenke, F., Poole, B., Ganguli, S., 2017. Continual learning through synaptic intelligence. arXiv: 1703.04200 .

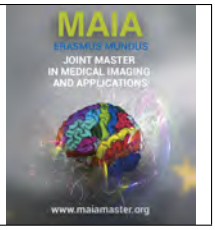
Zhang, H., Valcarcel, A.M., Bakshi, R., Chu, R., Bagnato, F., Shinohara, R.T., Hett, K., Oguz, I., 2019. Multiple sclerosis lesion segmentation with tiramisu and 2.5d stacked slices, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Springer International Publishing. pp. 338–346. doi:10.1007/978-3-030-32248-9_38.

Zhang, H., Zhang, J., Li, C., Sweeney, E.M., Spincemaille, P., Nguyen, T.D., Gauthier, S.A., Wang, Y., Marcille, M., 2021. All-net: Anatomical information lesion-wise loss function integrated into neural network for multiple sclerosis lesion segmentation. NeuroImage: Clinical 32, 102854. doi:10.1016/j.nicl.2021.102854.

Zhang, J., Gu, R., Xue, P., Liu, M., Zheng, H., Zheng, Y., Ma, L., Wang, G., Gu, L., 2023. S3r: Shape and semantics-based selective regularization for explainable continual segmentation across multiple sites. IEEE Transactions on Medical Imaging 42, 2539–2551. doi:10.1109/TMI.2023.3260974.

Medical Imaging and Applications

Master Thesis, June 2024




# Interpretable Lung Nodule Archetypes for Malignancy Classification

Muhammad Zain Amin[1,2], Joseph Y. Lo[1]

[1] *Center for Virtual Imaging Trials, Carl E. Ravin Advanced Imaging Laboratories, Department of Radiology, Duke University School of Medicine, Durham, NC, United States*
[2] *University of Girona, Spain*

**Abstract**

*Background:* Lung cancer screening relies heavily on accurately classifying the lung nodules malignancy. Current AI solutions, despite their high accuracy, often face skepticism from radiologists due to lack of interpretability and uncertainty associated with malignancy classification. To bridge this gap, we propose an interpretable approach using four distinct "archetypes" of lung nodules—Size, Spiculations, Lobulations, and Attachments—each contributing to the malignancy classification process. This study estimates the performance of interpretable archetypes against traditional radiomics features for malignancy classification. *Methods:* Our approach computed the archetypes using area distortion metric from angle-preserving spherical parameterization and used region growing concepts for accurate lung nodule, and vessel/wall attachment segmentation. We then integrated the archetypes into decision tree and neural network models to provide transparent and understandable classification results. The contribution of each archetype to malignancy prediction was quantified from SHAP (SHapley Additive exPlanations) values. We conducted a comprehensive evaluation using multiple datasets, including the LIDC, LungX, and Private Duke Lung Cancer Screening Dataset. Additionally, we incorporated uncertainty estimations to assess the reliability of classification results, comparing models based on interpretable archetypes against those based on traditional radiomics features. *Results:* By using archetypes, we achieved better AUC scores as compared to radiomics. Our models were trained on weakly-labeled radiological malignancy (LIDC_RM) cases and tested on strongly labeled pathological malignancy (LIDC_PM) cases, as well as externally available LungX and Duke cases. Uncertainty estimations also indicated that the archetypes-based neural network model provided more reliable results compared to the radiomics-based neural network model. *Conclusion:* Our proposed approach offers a reliable and transparent tool for classifying lung nodule malignancy, which has the potential to foster greater trust among radiologists. By effectively balancing accuracy and interpretability, the archetypes approach holds promise for enhancing clinician confidence in AI-assisted diagnoses. Incorporating uncertainty estimations further confirms the model's reliability, making it a valuable asset in clinical settings.

*Keywords:* Spiculations, Radiomics, Interpretable AI, Lung Cancer Screening

## 1. Introduction

Over recent years, artificial intelligence (AI) solutions have significantly impacted both industry and academia, revolutionizing radiologic research. Numerous studies have employed AI techniques to address complex medical imaging challenges (Hosny et al., 2018), often demonstrating that AI models can surpass human radiologists in performance (McKinney et al., 2020) (Hou et al., 2021) (Ardila et al., 2019). As a result, an increasing number of AI-driven products have been developed for clinical applications (van Leeuwen et al., 2021), profoundly influencing clinicians' daily diagnostic practices.

A significant challenge hindering the widespread adoption of AI in critical decision-making areas like healthcare is the lack of model interpretability and transparency (Rudin, 2019) (Reyes et al., 2020) (Caspers, 2021). It is essential to develop AI models that are both reliable and explainable for clinical use. For AI to serve as a valuable second opinion for radiologists, it must

provide diagnostic results that are both accurate and understandable. Addressing the "black box" nature of AI has become a key research priority, focusing on two main approaches to interpreting AI models: the "why" approach, which seeks to explain the rationale behind diagnostic outcomes, and the "where" approach, which aims to visualize the key regions the AI considers important.

(Ribeiro et al., 2016) introduced the "LIME" method, which aims to explain AI models by pinpointing informative segments within the input data. This approach has found success in natural language processing (NLP) and has been adapted for use in computer vision. Similarly, (Ghorbani et al., 2019) proposed a concept-based explanation technique that seeks to identify human-interpretable regions within input images. Furthermore, instance-based methods have been developed to uncover relationships between training data and test data by locating similar examples within the training set (Chen et al., 2019).

Unlike the previously mentioned methods that focus on explaining AI model predictions, class activation mapping (CAM) techniques (Zhou et al., 2016) (Selvaraju et al., 2017) create saliency maps to highlight the regions that the AI model activates. Some CAM-based methods have been utilized in various healthcare applications, including COVID-19 diagnosis (Oh et al., 2020), skin lesion detection (Zhang et al., 2019), determining the lymph node status in early-stage breast cancer (Zheng et al., 2020), and predicting extracapsular extension in prostate cancer (Hou et al., 2021). However, CAM-based methods have several limitations. They struggle with precise localization of the regions contributing to the decision, especially in complex images where features are subtle or overlapping (Arun et al., 2021). Additionally, while CAMs provide a visual explanation of what the model focuses on, the explanations can be coarse and may not always be easy for clinicians to interpret. Furthermore, CAMs cannot not generalize well across different datasets, and they primarily offer qualitative insights without quantitative measures of feature importance, which limit their use in clinical settings (Rudin, 2019).

Lung cancer is the leading cause of cancer-related deaths worldwide (Bade and Cruz, 2020). For many years, computer-aided diagnosis (CAD) models have been developed to assist in the evaluation of pulmonary nodules (Zhao et al., 2012) (Cao et al., 2020). These CAD systems primarily focus on two tasks: the detection of nodules and their classification. Historically, most research has prioritized improving the accuracy of these models, often at the expense of model interpretability (Xie et al., 2018) (Xie et al., 2019). CAD systems are intended to support clinicians by providing a second opinion, aiding in accurate diagnoses. However, the lack of interpretability in these models poses a significant challenge, as it hinders radiologists' ability to validate or refute the CAD system's predictions. This lack of transparency is a major barrier to the integration of AI models into everyday clinical practice. In recent years, researchers have increasingly recognized the importance of interpretability in AI models. Several studies have focused on creating explainable diagnostic models for pulmonary nodules by providing predicted clinical characteristics, such as calcification, sphericity, and subtlety (Shen et al., 2019) (Liu et al., 2019).

In recent years, when evaluating lung nodules for malignancy, specific archetypes such as the size of the nodules, number of spiculations, number of lobulations, and number of attachments are proving critical according to guidelines by organizations like the Food and Drug Administration (FDA) and the Fleischner Society. These archetypes are considered interpretable as they provide clear, understandable criteria that can be used to assess the likelihood of malignancy. For instance, lung nodules smaller than 6 mm generally have a very low risk of malignancy typically less than 1 percent, but the risk increases with size. Nodules measuring 6-8 mm in high-risk patients require follow-up, and those larger than 8 mm may need immediate further diagnostic procedures like CT scans, PET/CT, or biopsy (Gould et al., 2015) (MacMahon et al., 2017). The presence of spiculations, which indicates spiky or irregular margins, also significantly raises the suspicion of cancer and warrants further investigation regardless of nodule sizes. Similarly, lobulated nodules with uneven or lobed margins suggest abnormal growth patterns often associated with cancer (Larici et al., 2017). Nodules attached to the vessels/walls of the lung or pleura are also considered more suspicious, particularly when combined with other malignancy archetypes such as spiculations and lobulations, requiring careful evaluation and follow-up (MacMahon et al., 2017). These archetypes are not only crucial for assessing malignancy but are also considered interpretable because they can be directly observed, measured, and evaluated in a consistent manner.

In our proposed work, we present a comprehensive pipeline to measure and utilize interpretable archetypes of nodules such as size, lobulations, spiculations, and vessel/wall attachments, and use them for the malignancy classification purpose.

The contributions of our research are as follows:

1. Key interpretable lung nodule archetypes for malignancy estimation are computed by using area distortion metric from angle-preserving spherical parameterization technique.
2. Region growing segmentation concepts are used for accurate lung nodule segmentation, and vessel/wall attachment segmentation. This approach leads to accurate spiculation quantification because it helps to exclude the attachments from spiculations present on the lung nodule surface.
3. Moreover, a malignancy classification model is

introduced, which only uses four interpretable archetypes (size, spiculations, lobulations, vessel/wall attachments) and has the advantage of utilizing weak-labeled training data. These archetypes are integrated into a decision tree and neural network for malignancy classification. Also, the contribution of each archetype to the malignancy is calculated based on SHAP values, which use game theory to assign credit for malignancy prediction to each archetype.

4. Monte Carlo dropout uncertainty estimations were incorporated to assess the neural network's reliability. The uncertainty estimation trends showed that the archetype-based neural network model provided more reliable classification results than the radiomics-based neural network model.

## 2. State of the art

For malignancy classification, radiomics, which involves the extraction of a large number of quantitative features from medical images, can indeed be challenging to interpret. This complexity arises from the high dimensionality and abstract nature of the features extracted, which often lack direct and intuitive clinical meaning. Many radiomics features include higher-order statistical measures, texture patterns, and wavelet transforms, which can be difficult for clinicians to understand without a clear context of their relevance to clinical outcomes. Additionally, the lack of standardization in feature extraction and reporting across studies further complicates the interpretability of radiomics data. While some radiomics features show promise in correlating with clinical outcomes like tumor grade or patient prognosis, many lack direct clinical correlation, making it challenging to translate these features into actionable insights.

Recent advancements in radiomics have enabled the application of this technology in various clinical settings (Hawkins et al., 2016). By extracting a multitude of quantitative features from medical images and employing data mining techniques, radiomics studies can predict tumor responses and patient outcomes with enhanced precision. This approach has led to more accurate predictions of local tumor control and overall patient survival. For an in-depth review of radiomics and radiogenomics research focused on forecasting clinical outcomes in lung cancer, you can refer to the comprehensive study (Thawani et al., 2018).

Radiomics analysis for lung cancer screening has been investigated by several researchers. For instance, (Hawkins et al., 2016) developed a random forest classifier that utilized 23 radiomics features for malignancy classification. In a similar vein, (Buty et al., 2016) proposed another random forest classifier, which employed spherical harmonics feature extractor for 400 shape features, and also utilized a pre-trained deep neural network-based feature extractor for 4096 appearance features. The spherical harmonics technique provide a decomposition of frequency-space basis for representing all the functions defined over sphere. They are essential for accurately describing the overall shape of an object. Moreover, while effective for describing the general shape, spherical harmonics are less capable of capturing localized features, such as spiculations, that are specific to certain regions of a shape.

In another study, (Kumar et al., 2017) introduced a deep neural network model that leveraged five thousand distinct features for the detailed analysis. (Liu et al., 2017) also proposed a linear classifier model based on twenty-four image characteristics that were visually recorded by the physicians. Also, (Choi et al., 2018) created a malignancy classification framework for lung nodules by utilizing the support vector machine classifier combined with a selection operator and, a least absolute shrinkage, by employing just two radiomics features, texture and, size. Although these radiomics studies have enhanced classification accuracy, they are limited by the lack of clinical and biological interpretability.

Efforts to improve interpretability include feature reduction techniques, robust clinical validation, and advanced visualization tools to better understand the spatial distribution and relevance of radiomics features within anatomical and pathological contexts. Despite its potential to enhance diagnostic and prognostic accuracy, improving the interpretability of radiomics remains a critical area of ongoing research (Lambin et al., 2017) (Zwanenburg et al., 2020).

The edge characteristics of lung nodules, such as spiculations, which are spike-like projections on the nodule's surface, can play a crucial role in assessing malignancy risk (Swensen et al., 1997). Malignant nodules often exhibit irregular and blurred boundaries, in contrast to benign nodules, which tend to have well-defined and smooth edges. To streamline lung cancer screening using CT images, the American College of Radiology developed the Lung Imaging Reporting and Data System (Lung-RADS). This system standardizes lung cancer assessments based on nodule size, appearances (including lobulations, spiculations and vessel/wall attachments), as well as calcification (McKee et al., 2016). According to Lung-RADS, the presence of spiculation is a significant indicator that increases the likelihood of malignancy, thereby enhancing the accuracy of predictions.

Quantifying spiculations in pulmonary nodules has been explored, though not primarily for malignancy prediction. (Niehaus et al., 2015) developed a computer-aided diagnosis (CAD) system that utilized shape features dependent on nodule size to measure spiculations. Similarly, (Ciompi et al., 2015) introduced a frequency-based shape descriptor aimed at evaluating spiculations

in detected nodules for the lung cancer screening purpose. (Dhara et al., 2016) concentrated on quantifying spiculations on a surface mesh derived from binary mask of segmented nodules. They employed geodesic distance transformation and mean curvature to detect spiculations and identified the baseline by tracking sudden surface changes. However, their method's accuracy was compromised by its sensitivity to local surface variations, making it challenging to precisely identify the baseline for noisy spiculation peaks.

## 3. Material and methods

The first part of our proposed methodology focuses on quantifying sharp spikes present on lung nodules in an interpretable way. The visual steps have been shown in *Figure 1*.

### *3.1. Quantifying Spikes Pipeline*

We use the conformal spherical parameterization to map the shape of a nodule onto a sphere while preserving angles, enabling accurate measurement of the spikes. In our case, the conformal spherical parameterization involves several steps in the following order: -

(i) *Nodule Surface Division:*

We take the shape of the lung nodule as a three-dimensional surface $S$. The surface $S$ is divided into two disk-like parts by calculating the eigenfunction of the Laplace-Beltrami operator, $\Delta_B$, which helps identify a boundary $\Gamma$ that splits $S$ into two parts, $S_1$ and $S_2$.

$$\Delta_B \phi = \lambda \phi$$

Here, $\Delta_B$ is the Laplace-Beltrami operator, $\phi$ is the eigenfunction, and $\lambda$ is the eigenvalue. The boundary $\Gamma$ is defined where the eigenfunction $\phi$ takes a specific value, often zero (the nodal line):

$$\Gamma = \{x \in S \mid \phi(x) = 0\}$$

The surface $S$ is then divided into:

$$S_1 = \{x \in S \mid \phi(x) > 0\}$$

$$S_2 = \{x \in S \mid \phi(x) < 0\}$$

(ii) *Spherical Mapping:*
- *Flattening the Disk-like Parts:* After dividing the surface, each disk-like part is flattened using the Ricci flow algorithm, which ensures that the angles of the original nodule shape are preserved during the flattening process. The Ricci flow is described by the following equation:

$$\frac{\partial g_{ij}}{\partial t} = -2R_{ij}$$

where $g_{ij}$ is the metric tensor, $t$ is the time parameter, and $R_{ij}$ is the Ricci curvature tensor.

- *Projecting the Flattened Parts onto a Sphere*: Once the parts are flattened, they are reassembled into a spherical shape through a process called stereographic projection. The stereographic projection maps the flat parts onto a sphere and is given by:

$$x = \frac{2u}{1+u^2+v^2}$$
$$y = \frac{2v}{1+u^2+v^2}$$
$$z = \frac{1-u^2-v^2}{1+u^2+v^2}$$

where $(u, v)$ are coordinates in the plane and $(x, y, z)$ are coordinates on the sphere.

(iii) *Spikes Detection:*
In the process of mapping the lung nodule surface onto a sphere, the areas of the nodule surface get distorted. Area distortion refers to the changes in the size of different regions of the nodule surface as they are transformed to fit onto the sphere. We measure this distortion because when a region on the nodule surface shrinks (i.e. negative distortion), it indicates the presence of spikes. This shrinkage is important because it precisely characterizes the sharp spikes on the nodule surface. To quantify the spikes, we measure the amount of distortion at each point on the nodule surface using the following equation.

$$\epsilon_i := \log \frac{\sum_{j,k} A([\phi(v_i), \phi(v_j), \phi(v_k)])}{\sum_{j,k} A([v_i, v_j, v_k])}$$

where, $\epsilon_i$ calculates the logarithm of the ratio of the summed areas in the transformed space $\phi(v_i)$ to the original space. This ratio indicates how much the area has changed during the transformation. The detection process involves these steps:

- *Baseline and Apex Detection:* We identify baseline points, where the area distortion is zero, and apex points, where the area distortion reaches its maximum, indicating the sharpest spikes.

- *Height and Width Calculation*: For each detected spike, we calculate its height (the vertical extent of the spike) and width (the horizontal extent at half the height). This method allows us to precisely characterize the spikes on the lung nodule surface by analyzing the area distortion patterns.

By successfully applying the *Quantifying Spikes Pipeline*, we detected the spikes present on the nodule

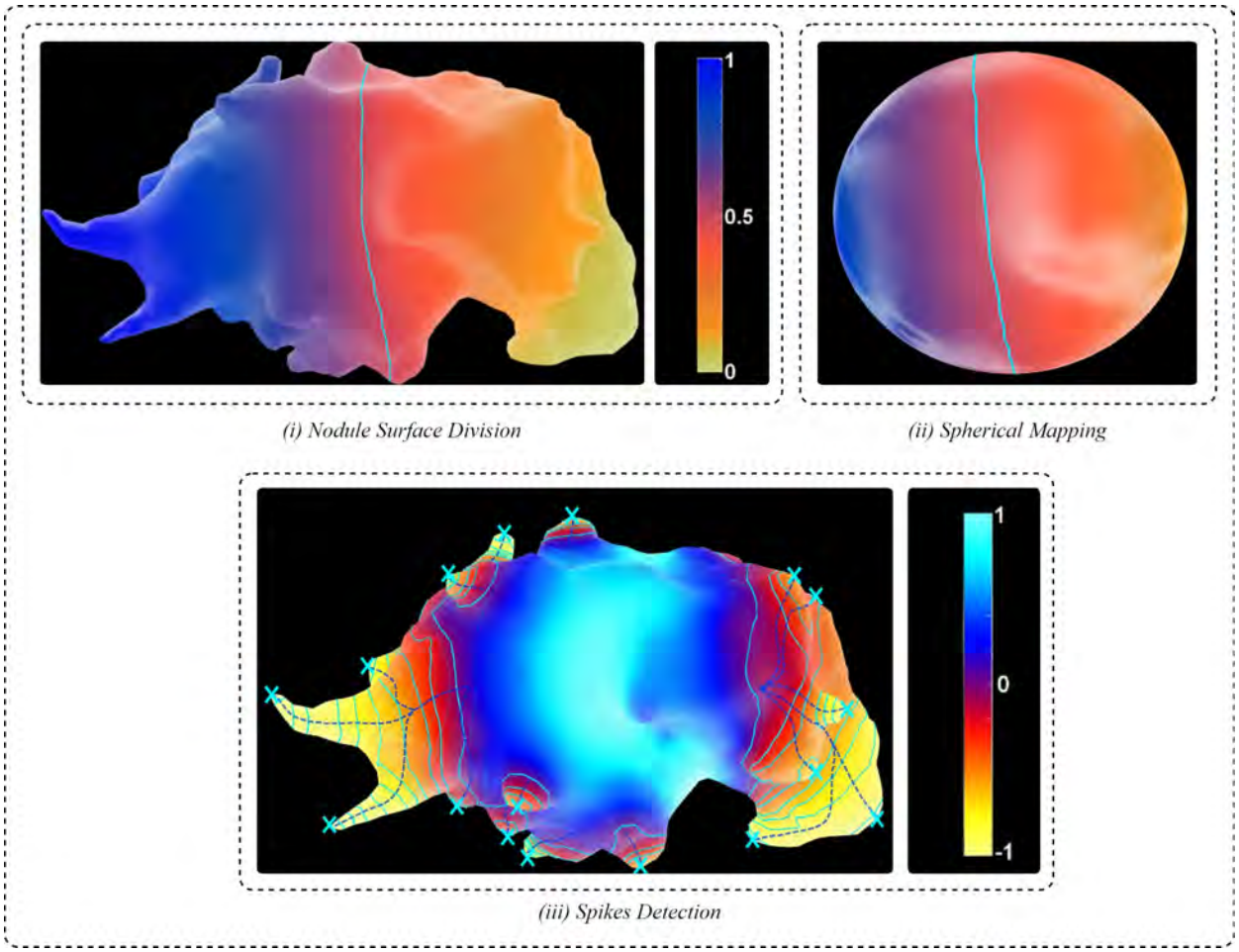



Figure 1: Quantifying Spikes Pipeline (i) *Nodule Surface Division:* The first significant eigen function of the Laplace-Beltrami operator is computed for the nodule's surface (mesh). This step identifies a boundary (light blue curve) that divides the nodule's surface into two halves. (ii) *Spherical Mapping:* Each half is flattened and then mapped onto a sphere while preserves angles. (c) *Spikes Detection:* The area distortion metric is used here. Areas with significant shrinkage (light blue Xs) indicate the highest points of spikes. The dark blue curves measure the height of these spikes.

surface. As we know, malignant lung nodules often exhibit irregular, lobulated, or spiculated margins due to the invasion of malignant cells into the pulmonary interstitium. Thus, our *Quantifying Spikes Pipeline* can identify both lobulations and spiculations. By classifying detected spikes into spiculations (sharp peaks) and lobulations (curved peaks), we enhance the archetypes for malignancy classification. To differentiate the lobulations from spiculations, we use a thresholding approach based on height threshold ($Th \geq 3$ mm) and solid angle ($T\Omega \leq 0.65$ sr). The height and solid angle thresholds were recommended in (Dhara et al., 2016). Additionally, we employed the full width at half maximum (FWHM) concept to achieve more accurate width measurements, considering the surface and its area distortion.

### *3.2. Region Growing Segmentation*

We employed region growing segmentation methods to precisely preserve spikes and segment the vessel/wall attachments of the nodules. For each nodule, a consensus contour was created by merging two or more contours utilizing the simultaneous truth and performance level estimation method (Warfield et al., 2004) (Choi et al., 2016), which served as our ground truth. While numerous region growing segmentation methods exist for nodule segmentation, most focus on the core regions of the nodule and are often complex to implement. To ensure reliability in segmentation, we incorporate two well known and established straightforward techniques: *Chest Imaging Platform* (CIP) segmentation (Yip et al., 2017) and the *GrowCut* (Vezhnevets and Konouchine, 2005) segmentation.

The GrowCut method is a region growing segmentation technique that utilizes two sets of seed points: one for the foreground and the other one for the background. These seed points compete to expand their respective regions until the algorithm converges. GrowCut is known for its simplicity and effectiveness in segmenting complex structures. One of its main advantages is its robustness in handling noisy images, which is particularly

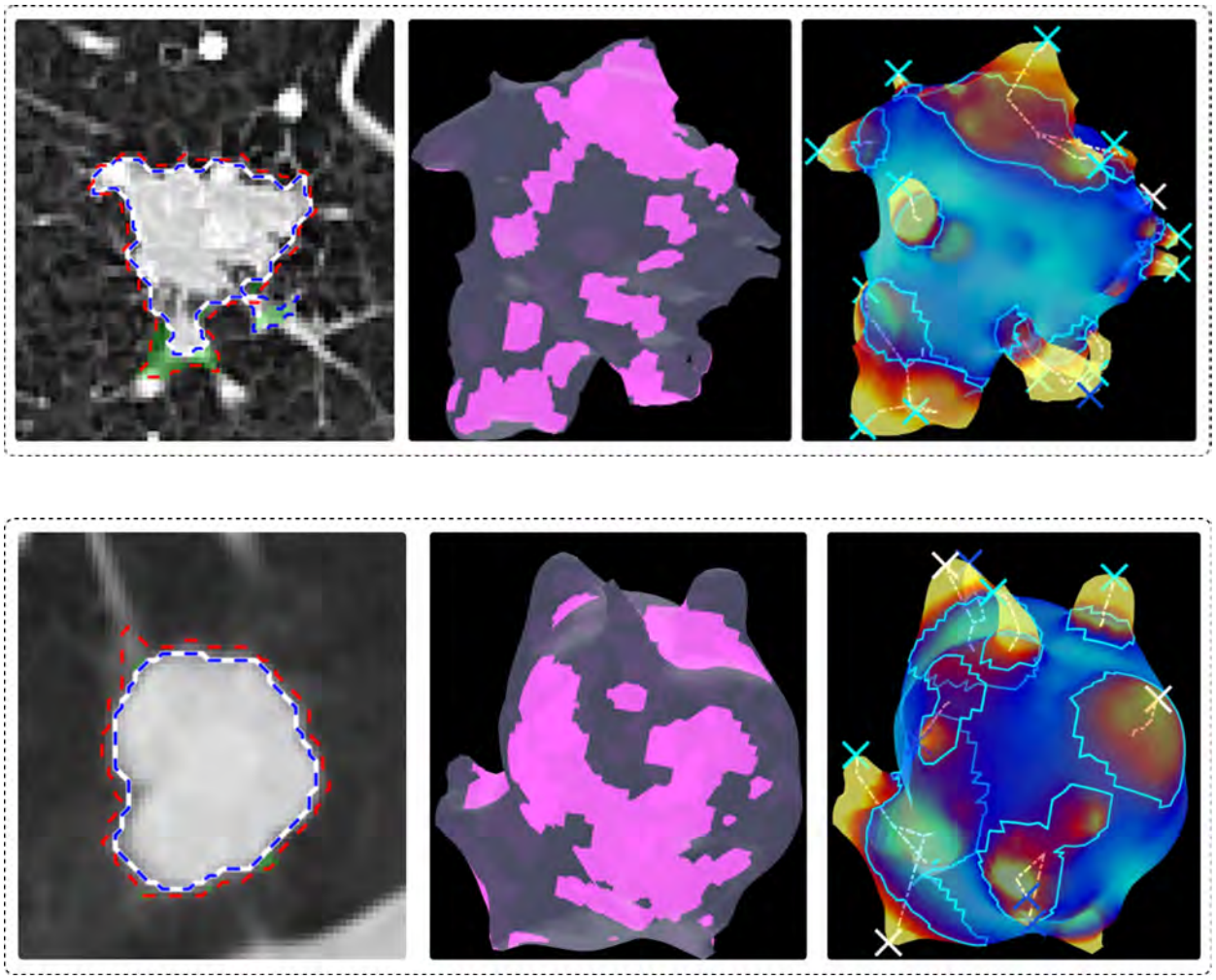


Figure 2: Region-growing segmentation and vessel/wall attachments detection. (*Left Images*) Segmentation results on an axial slice (red dashed line: CIP segmentation, blue dashed line: GrowCut segmentation, and white line: final segmentation). (*Middle Images*) 3D shapes of final segmentation and attachment regions (pink region: attachments). (*Right Images*) Spikes classification results (blue lines: baselines, green Xs: spiculations, white Xs: lobulations, blue Xs: attachments).
*Case 1 results explanation: Archetypes* classification results. (number of spiculations: 14, number of lobulations: 1, number of attachments: 9).
*Case 2 results explanation: Archetypes* classification results. (number of spiculations: 4, number of lobulations: 3, number of attachments: 5).

useful for medical image segmentation where noise can be an issue. Additionally, GrowCut is quite simple to implement and only requires minimal parameter tuning, making it easily accessible for various applications. The method uses iterative updates to propagate the seeds, ensuring detailed segmentation. However, due to its aggressive region-growing nature, it can sometimes extend into surrounding areas, such as the airway walls, chest wall, and vessel-like structures. Despite this, its ability to produce detailed and accurate segmentations with minimal user input makes it a valuable tool in medical imaging (Vezhnevets and Konouchine, 2005).

The CIP method is the level set-based segmentation algorithm that employs a front propagation strategy, starting from the seed point placed within the lung nodule. This segmentation approach, is guided by the feature maps of surrounding structures in order to prevent leakage into the adjacent areas. However, CIP method can sometimes miss parts of the tumor due to inaccurate wall and vessel feature maps. To overcome these limitations and leverage the strengths of both methods, we combined the GrowCut and CIP segmentation techniques. GrowCut's robustness in noisy environments and aggressive region-growing capabilities complement CIP's precise front propagation, resulting in improved vessel/wall attachments detection. This hybrid approach helps to mitigate the individual shortcomings of each method. Both GrowCut and CIP segmentation methods are publicly available in the 3D Slicer platform (Fedorov et al., 2012), making them accessible for research and clinical applications. CIP's advantage lies in its ability to accurately follow the nodule boundaries when the feature maps are reliable, while GrowCut provides a more robust segmentation in the presence of noise and complex anatomical structures. By integrating these methods, we aim to achieve more accurate and

reliable nodule segmentation (Yip et al., 2017).

#### 3.2.1. Radiomics Features

Following the implementation of our *Region Growing Segmentation Pipeline* to segment lung nodules, we also extracted a comprehensive set of 103 radiomics features from each lung nodule to quantitatively assess its intensity, texture, and shape characteristics (Choi et al., 2018):

(i) *Intensity features*
These first-order statistical measures capture the distribution and levels of CT attenuations within a nodule. Key intensity features include Minimum, Mean, Median, Maximum, Standard Deviation, Kurtosis, and Skewness, providing a detailed profile of the nodule's intensity variations.

(ii) *Shape features*
These features describe the geometric properties of the nodules. We measured the diameter, volumn, roundness, elongation, and flatness, which collectively characterize the three-dimensional structure and form of the nodules.

(iii) *Texture features*
To quantify the tissue density patterns, we employed advanced texture analysis methods. Specifically, we employed the *Gray Level Co-occurrence Matrix* (GLCM) and *Gray Level Run Length Matrix* (GLRM) techniques. From the GLCM, we derived features such as Correlation, Inertia, Cluster Prominence, Entropy, Haralick's Correlation, Energy, Cluster Shade, and Inverse Difference Moment. From the GLRM, we extracted Gray-Level Non-uniformity, Short-Run Low Gray-Level Emphasis, Long-Run Emphasis, High Gray-Level Run Emphasis, Long-Run High Gray-Level Emphasis, Low Gray-Level Run Emphasis, Short-Run High Gray-Level Emphasis, Long-Run Low Gray-Level Emphasis, Short-Run Emphasis, Run Length Non-uniformity. To ensure rotational invariance of these texture features, we computed their mean and standard deviation across 13 different directions.

### 3.3. Malignancy Classification Model

For malignancy classification, we evaluated our interpretable archetypes against the uninterpretable radiomics features extracted from lung nodules. To compare models' performances, we exclusively used archetypes measures, and integrated them into both decision tree and neural network models for predictions. After making prediction, we calculated the contribution of each archetype for the malignancy prediction using SHAP values, thus giving prediction insights made by the model to the radiologists.

#### 3.3.1. SHAP values

SHAP values are a concept from cooperative game theory to fairly distribute the total gains to all players based on their contribution to the total outcome. In our prediction model, SHAP values are used to explain the output of the prediction model by attributing the contribution of each archetype to the final prediction. SHAP values ensure fairness by distributing the contribution of each archetype based on its marginal contribution to different subsets of archetypes. Each archetype is considered a player in a coalition, and the SHAP value represents the average contribution of an archetype over all possible coalitions. The concept of marginal contribution is essential, as it quantifies the value added by including an archetype in a coalition of other archetypes (Lundberg and Lee, 2017).

Here is the detailed explanation to calculate the SHAP values of each archetype towards the final prediction.

- *Permutations of Archetypes:* Consider all possible permutations of the archetypes.

- *Marginal Contribution:* For each permutation, compute the marginal contribution of an archetype by comparing the model's output with and without the archetype.

- *Average Contribution:* Average the marginal contributions of the archetype over all permutations.

For an archetype $i$, the Shapley value $\phi_i$ is calculated as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! \cdot (|N| - |S| - 1)!}{|N|!} \left(v(S \cup \{i\}) - v(S)\right)$$

where,

- $N$ is the set of all archetypes
- $S$ is the subset of archetypes excluding $i$
- $|S|$ is the number of archetypes in subset $S$
- $v(S)$ is the value (e.g., prediction) of the subset $S$
- $v(S \cup \{i\})$ is the value of the subset $S$ with archetype $i$ included

In parallel, we employed radiomics features, integrating them into decision tree and deep neural network models to assess their overall performance compared to the archetype-based predictions. Prior to feeding the radiomics features into these models, we performed feature selection using Recursive Feature Elimination (RFE) with a Logistic Regression estimator. The RFE method identifies the most relevant features by recursively removing the least important ones and fitting the model multiple times. Through this process, we selected 10 features that contribute

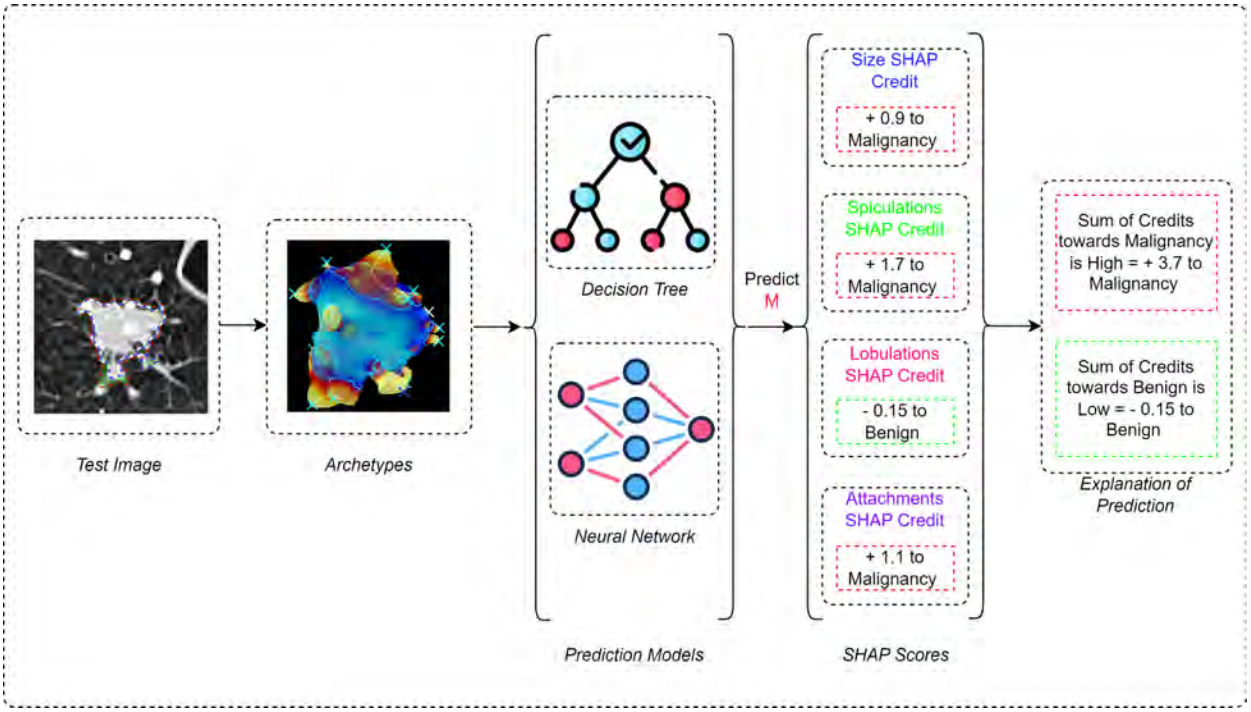



Figure 3: *Malignancy Prediction Assessment Flowchart:* Pipeline illustrating the process of lung nodule malignancy prediction assessment and explanation. The workflow begins with a test image of a lung nodule, which is processed to measure archetypes such as size of the nodule, number of spiculations, number of lobulations, and number of attachments through *Quantifying Spikes Pipeline*, *Region Growing Segmentation*, and (*geometric property of volume for the size measurement*). These archetypes are fed into two prediction models: a decision tree and a neural network for the prediction. After prediction on the test image, we measure SHAP credit of each archetype. Positive SHAPs values credits towards malignancy (+ sign is the malignancy class magnitude) and negative SHAPs values credits towards benign (- sign is the benign class magnitude). *For the same test image case use here in Figure 3, we already calculated the Archetypes Measures (Refer to Case 1 of Figure 2)*: Size of Nodule = 6.8 mm, number of spiculations =14, number of lobulations = 1, number of attachments =9
*Results of Explanation:* For this test case, the malignancy class has an aggregated SHAP credit of + 3.7, which is significantly higher than the benign class aggregate SHAP credit of - 0.15. This indicates that the prediction made by the model also considers a strong influence from the nodule size of 6.8 mm (considered high risk), the number of spiculations (14 – considered high), and the number of attachments (9 – considered high) in predicting the test case as malignant. This whole assessment can help the radiologists whether they should consider the prediction made by the model by looking at the calculated archetypes measures and then the prediction made by the model on the test case along with quantified SHAP scores showing the importance of each archetype in the final decision made by the model.

the most to the classification task: *BoundingBoxSize1*, *BoundingBoxSize2*, *EquivalentSphericalPerimeter*, *MeanOfLongRunEmphasis*, *OrientedBoundingBoxSize1*, *OrientedBoundingBoxSize3*, *PrincipalMoments1*, *StandardDeviationOfGreyLevelNonuniformity*, *StandardDeviationOfLongRunEmphasis*, and *StandardDeviationOfRunLengthNonuniformity*. These features were found to be the most significant in classifying malignancy, providing a robust subset for further analysis but lack clinical interpretation to understand predictions.

### *3.3.2. Uncertainty Estimations*

Uncertainty estimations were also incorporated in order to assess the neural network's prediction's reliability. Uncertainty estimation in neural networks helps quantify the confidence in the predictions made by the model. This is particularly important for our proposed prediction model, where understanding the reliability of the model's predictions is also very important. We have used the Monte Carlo (MC) Dropout technique to estimate model uncertainty (epistemic uncertainty) in neural networks (Gal and Ghahramani, 2016). It involves using dropouts during both training and inference phases to approximate the uncertainty.

Given below is the mechanism of MC Dropout technique we used to estimate the uncertainty estimations:-

(i) *Dropout during training*
- Dropout is a regularization technique where, during each training iteration, a fraction of the neurons are randomly dropped (i.e., set to zero). This prevents the network from overfitting and encourages it to learn robust features.
- In our neural network prediction model, dropout is applied with a probability of 0.5.

(ii) *Dropout during inference*
- During inference, dropout is usually turned off to use the full network for predictions. However, for Monte Carlo Simulations, dropout is kept on, and multiple stochastic forward passes are performed.

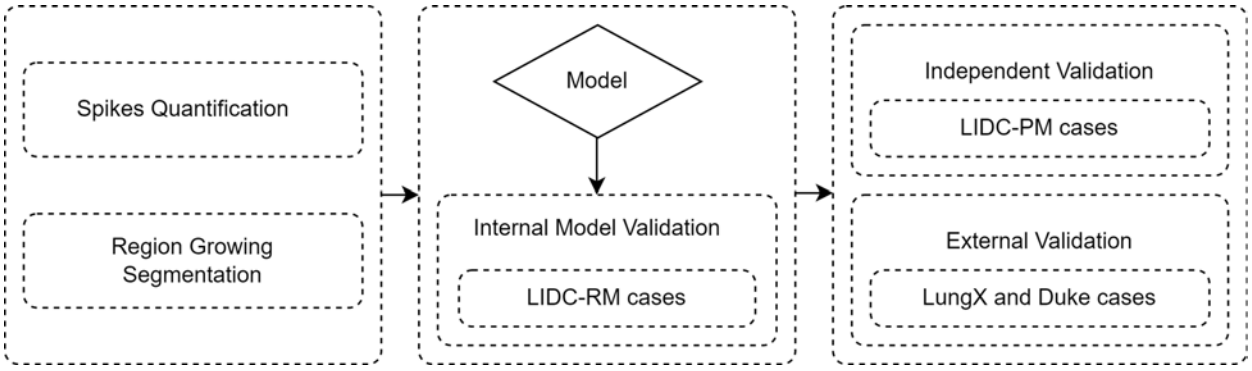



Figure 4: Prediction Model Evaluation Criteria

• Each forward pass uses a different random subset of the network, simulating the effect of sampling from an ensemble of models.

(iii) *Predictive Distribution*
• By performing multiple forward passes (in our case: 50), we obtain a distribution of predictions for each input.
• The mean of these predictions gives the final predicted output.
• The standard deviation of these predictions provides an estimate of the uncertainty.

## 4. Results

### *4.1. Data Preparation*

The Lung Image Database Consortium dataset (LIDC) (Armato III et al., 2011), LungX dataset (Armato III et al., 2016), and a few samples from the Private Duke Lung Cancer Screening dataset were used to evaluate our proposed methodology.

The LIDC dataset contains 1,018 cases with low-dose screening thoracic Computed Tomography scans and annotated lesions. Four experienced radiologists annotated nodules. Out of the entire dataset, 883 cases included nodules with contour annotations. For each case, we employed our region growing segmentation approach to segment lung nodules while preserving spikes present on the nodule surface. The accuracy of our region growing segmentation approach was measured by the Dice coefficient, resulting in a score of $0.746 \pm 0.11$.

The LungX dataset includes 83 strongly-labeled pathological malignancy cases. Additionally, we utilized 335 cases from the Duke Lung Cancer Screening dataset. We applied the same region-growing segmentation approach to the nodules in each case of these datasets to accurately segment lung nodules and detect vessel/wall attachment regions.

For a detailed performance analysis of our classification model, we used two different subsets of the LIDC dataset: a strongly-labeled subset of pathological malignancy cases LIDC_PM and a weakly-labeled subset of radiological malignancy cases LIDC_RM. We further subdivided the LIDC_RM subset into two groups for model training and internal validation purposes.

For the independent validation of our classification model, we first validated the model on the strongly-labeled pathological malignancy cases LIDC_PM. Additionally, we performed external validation using the LungX and Duke cases as shown in the Figure 4.

### *4.2. Archetypes Distribution*

After implementing the *Quantifying Spikes Pipeline*, and *Region Growing Segmentation* on LIDC_RM cases, we were able to compute the lung nodule archetypes such as the size of the nodule, number of spiculations, number of lobulations, and number of attachments before feeding them into the training models. We classified the spiculations and lobulations out of spikes based on height and solid angle thresholds. Using the same thresholds as in (Dhara et al., 2016), we clearly differentiated spiculations from lobulations and detected as many spiculations as possible without false positives. The final selected thresholds were height ($Th \geq 3$ mm) and solid angle ($T\Omega \leq 0.65$ sr).

The bar plots in *Figure 5* illustrate the percentage distribution of various lung nodule archetypes for both malignant and benign cases in the LIDC_RM dataset. The first plot, showing the percentage distribution of size, indicates that smaller nodules (less than 4mm) are predominantly benign with approximately 90% benign and 10% malignant. Nodules in the size range of 4mm-6mm show around 75% malignancy, while those in the 6mm-8mm range have more than 80% malignancy. Nodules in the 8mm-10mm range also reach around 80% malignancy. This trend suggests that nodule size is a significant indicator of malignancy.

The second plot, displaying the percentage distribution of the number of spiculations, reveals that benign nodules typically have zero or very few spiculations, with about 80% of nodules having 0-1 spiculations being benign and 20% being malignant. As the number of spiculations increases, the percentage of malignant nodules also increases significantly. For nodules with

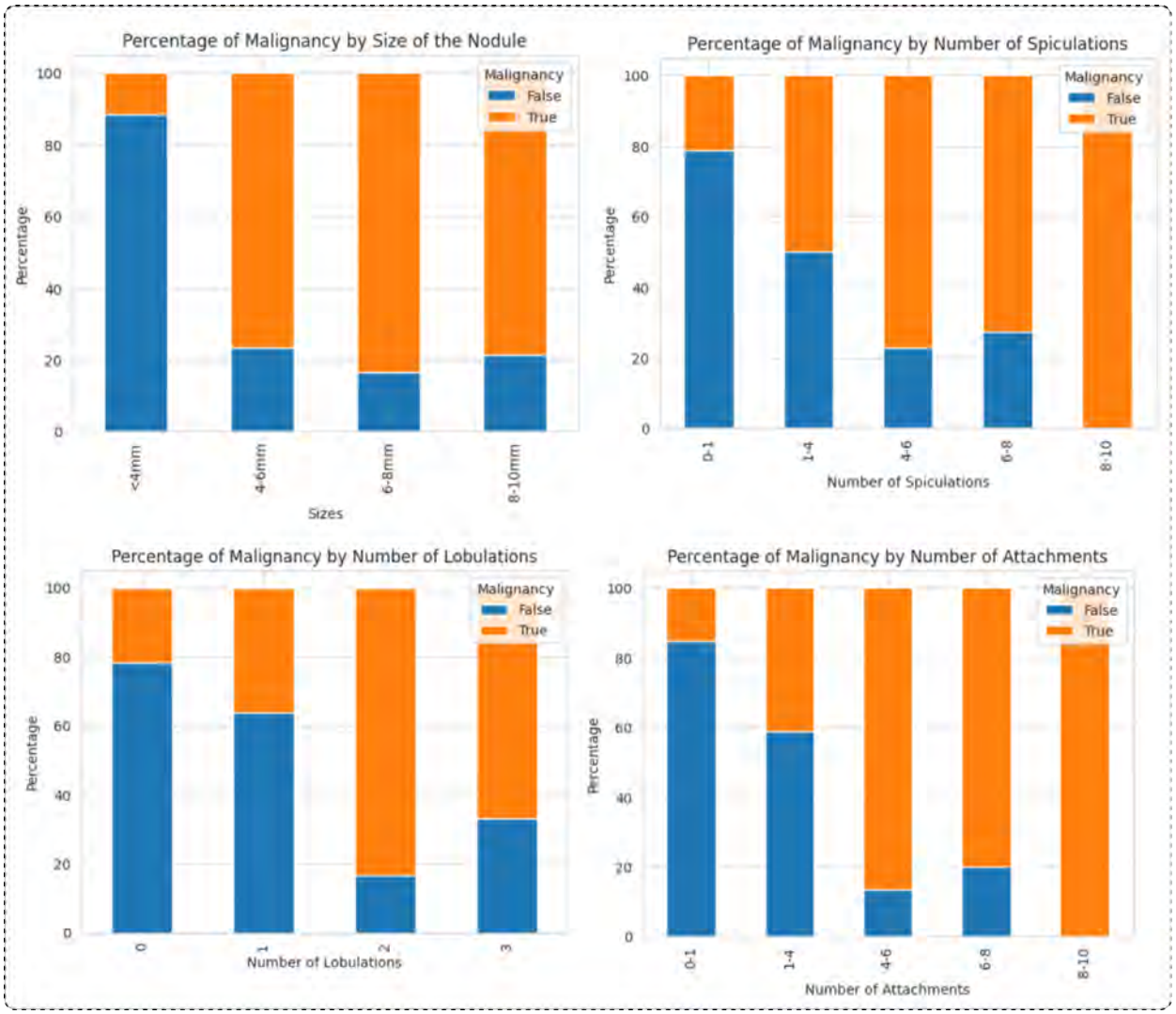



Figure 5: Bar plot results showing the percentage of malignant and benign cases for each lung nodule archetype (size, number of spiculations, number of lobulations, and number of attachments) across LIDC_RM cases. The x-axis represents different ranges and counts of the nodule archetypes, while the y-axis represents the percentage of nodules within each range or count. The blue bars represent benign nodules (Malignancy = False), and the orange bars represent malignant nodules (Malignancy = True).

1-4 spiculations, around 50% are malignant, while for those with 4-6 and 6-8 spiculations, the malignancy rate rises to about 80%. Nodules with 8-10 spiculations are almost entirely malignant. This indicates that the presence of multiple spiculations is a strong predictor of malignancy.

The third plot, showing the percentage distribution of the number of lobulations, indicates that benign nodules generally have zero lobulations, with about 80% being benign and 20% malignant. As the number of lobulations increases, the malignancy percentage also increases. Nodules with one lobulation have around 40% malignancy, those with two to three lobulations have about 70-80% malignancy. Although lobulations are less impactful than size and spiculations, their presence still increases the likelihood of malignancy.

Finally, the fourth plot, showing the percentage distribution of the number of attachments, indicates that benign nodules generally have zero or few attachments, with more than 80% of nodules having 0-1 attachments being benign and about 20% being malignant. For nodules with 1-4 attachments, the malignancy rate rises to about 40%, while for those with 4-6 and 6-8 attachments, the malignancy rate increases to approximately 80-85%. Nodules with 8-10 attachments shown almost sure chances of malignancy. This suggests that the presence of several attachments can also be an indicator of malignancy.

These bar plots provide a clear visualization of how different lung nodule archetypes are associated with malignancy, which is crucial for developing accurate and interpretable predictive models.

| Datasets | Archetypes based Decision Tree Model | | | Radiomics based Decision Tree Model | | |
|---|---|---|---|---|---|---|
| | AUC | Upper Bound | Lower Bound | AUC | Upper Bound | Lower Bound |
| LIDC-RM | 0.9192 | 0.9595 | 0.8648 | 0.8721 | 0.9347 | 0.8037 |
| LIDC-PM | 0.7286 | 0.8380 | 0.6318 | 0.7506 | 0.8560 | 0.6424 |
| LungX | 0.7085 | 0.8250 | 0.6608 | 0.6783 | 0.7946 | 0.5687 |
| DUKE | 0.8504 | 0.9102 | 0.7818 | 0.7734 | 0.8417 | 0.7064 |

Table 1: Comparison of Archetypes based and Radiomics based Decision Tree Model Results

| Datasets | Archetypes based Neural Network Model | | | Radiomics based Neural Network Model | | |
|---|---|---|---|---|---|---|
| | AUC | Upper Bound | Lower Bound | AUC | Upper Bound | Lower Bound |
| LIDC-RM | 0.9228 | 0.9626 | 0.8676 | 0.8918 | 0.9496 | 0.8052 |
| LIDC-PM | 0.7632 | 0.8566 | 0.7047 | 0.7419 | 0.8467 | 0.6234 |
| LungX | 0.7166 | 0.8267 | 0.6773 | 0.7538 | 0.8493 | 0.6422 |
| DUKE | 0.8511 | 0.9279 | 0.7834 | 0.7786 | 0.8621 | 0.6791 |

Table 2: Comparison of Archetypes based and Radiomics based Neural Network Model Results

### *4.3. Malignancy Classification Models*

(i) *Decision Tree:*

The decision tree model based on archetypes outperformed the radiomics-based model across various datasets. The archetype-based model achieved an AUC of 0.919 on LIDC_RM cases, 0.728 on LIDC_PM cases, 0.708 on LungX cases, and 0.850 on Duke cases. In comparison, the radiomics-based model showed lower AUC scores: 0.872 on LIDC_RM, 0.678 on LungX, and 0.773 on Duke cases. *Table 1.* illustrate highlighting the superior performance of the archetype-based decision tree model.

To quantify the reliability of these results, we calculated the confidence intervals for the AUC scores using bootstrap simulations. Bootstrap simulations involve repeatedly resampling the dataset with replacement to create numerous simulated samples. From these samples, we can estimate the distribution of the AUC scores. The confidence interval provides an upper and lower bound, indicating the range within which the true AUC score is likely to lie with a specified level of confidence (typically 95 percent ). For instance, the confidence interval for the archetype-based decision tree model on LIDC_RM cases ranged from 0.864 to 0.959. A narrower gap between the upper and lower bounds indicates more precise and reliable estimates. Conversely, a wider gap suggests greater variability and less certainty in the results. In our comparisons, the archetype-based model generally exhibited narrower confidence intervals compared to the radiomics-based model, indicating more reliable and consistent performance.

(ii) *Neural Network*

*Table 2* provides a comprehensive comparison of the performance of archetypes-based and radiomics-based neural network models across various datasets, highlighting their respective AUC scores and 95% confidence intervals. The archetypes-based neural network model consistently achieved higher AUC scores compared to the radiomics-based model for most datasets. For the LIDC-RM dataset, the archetypes-based model achieved an AUC of 0.922, with a confidence interval ranging from 0.867 to 0.962, whereas the radiomics-based model had an AUC of 0.891 with a wider confidence interval from 0.805 to 0.949. This narrower confidence interval of the archetypes-based model indicates more reliable and consistent performance.

In the LIDC-PM dataset, the archetypes-based model attained an AUC of 0.763, with a confidence interval from 0.704 to 0.856, compared to the radiomics-based model's AUC of 0.741 and confidence interval from 0.623 to 0.846. This further underscores the more precise and consistent performance of the archetypes-based model. Similarly, for the LungX dataset, although the radiomics-based model had a slightly higher AUC of 0.753 compared to the archetypes-based model's AUC of 0.716, the archetypes-based model's narrower confidence interval from 0.677 to 0.826 indicates greater reliability.

For the Duke dataset, the archetypes-based model achieved an AUC of 0.8511, with a confidence interval from 0.783 to 0.927, while the radiomics-based model had an AUC of 0.778 and a confidence interval from 0.679 to 0.862. Again, the archetypes-based model demonstrates a narrower confidence interval, highlighting its more reliable performance. Overall, the confidence intervals for the archetypes-based model are consistently narrower across all datasets, suggesting greater reliability and less variability in its performance. This indicates that the archetypes-based neural network

model not only outperforms the radiomics-based model in terms of AUC scores but also provides more consistent and reliable predictions, making it a more robust tool for lung nodule classification.

### *4.4. Uncertainty Estimations Results*

MC Dropout was used to estimate the uncertainty in neural network predictions by performing multiple stochastic forward passes with dropout enabled, resulting in a distribution of predictions for each input. The mean of these predictions gives the final output, while the standard deviation estimates the uncertainty. Lower standard deviation values indicate more consistent and reliable predictions, which is crucial for clinical decision-making. *Figure 6* show the cumulative mean uncertainty performance curves across different datasets, comparing neural network models using archetype features and radiomics features. For the LIDC-RM dataset, the archetype-based model starts with an uncertainty of 0.15, stabilizing around 0.05, while the radiomics-based model starts at 0.25, stabilizing at 0.15. In the LIDC-PM dataset, the archetype-based model stabilizes at 0.06, compared to 0.20 for the radiomics-based model. The LungX dataset shows the archetype-based model stabilizing at 0.05, while the radiomics-based model stabilizes between 0.20 to 0.25. For the Duke dataset, the archetype-based model stabilizes around 0.04, while the radiomics-based model stabilizes at 0.08. Across all datasets, the archetype-based model consistently demonstrates lower cumulative mean uncertainty, indicating more reliable and consistent predictions compared to the radiomics-based model. This highlights the archetype-based model's potential for improving clinical decision-making in lung nodule malignancy assessment by providing more dependable outputs.

## 5. Discussion

In this study, we introduced a novel approach for predicting lung nodule malignancy using interpretable lung nodule archetypes. Our results indicate that these archetypes provide a robust framework for malignancy prediction, outperforming traditional radiomics features in both accuracy and interpretability.

1. *Connection to Existing Literature*
   The need for interpretability in AI models, particularly in high-stakes fields like medical diagnostics, has been well-documented. Previous studies have highlighted the "black box" nature of many AI systems as a significant barrier to clinical adoption. Our approach directly addresses this issue by focusing on archetypes that are both clinically relevant and easily interpretable. For instance, the association between nodule size and malignancy risk is well-established, with larger nodules posing a higher risk. Similarly, spiculations and lobulations are recognized as critical indicators of malignancy, contributing to the suspicion of cancer. By leveraging these known indicators, our model not only enhances predictive accuracy but also provides clear, understandable reasons for its predictions.
2. *New Understandings*
   Our findings extend the current understanding of lung nodule assessment by demonstrating that a model based on interpretable archetypes can achieve superior performance compared to traditional radiomics-based models. This is particularly significant given the complexity and often abstract nature of radiomics features, which can be challenging for clinicians to interpret. The use of SHAP values to quantify the contribution of each archetype further enhances the transparency of our model, providing detailed insights to radiologists on how each archetype influences the prediction.
   The incorporation of uncertainty estimations using Monte Carlo Dropout is another critical advancement. Uncertainty quantification is crucial for clinical settings, where the confidence in a model's predictions can significantly impact decision-making. Our results show that the archetype-based models not only provide more accurate predictions but also exhibit lower uncertainty, making them more reliable for clinical use.
3. *Significance of the Results*
   The significance of our results lies in their potential to bridge the gap between AI model performance and clinical usability. By focusing on interpretable archetypes that clinicians are already familiar with, our model can foster greater trust and acceptance among radiologists. This is a crucial step toward the integration of AI-assisted diagnostics in routine clinical practice. Moreover, the improved performance and reduced uncertainty of our model highlight its potential for enhancing clinical decision-making and ultimately improving patient outcomes.

## 6. Conclusions

In conclusion, our interpretable approach to lung nodule malignancy prediction represents a significant advancement in the field of medical AI. By balancing accuracy with interpretability and incorporating uncertainty estimations, our approach provides a reliable and transparent tool for clinical decision-making, paving the way for greater adoption of AI in lung cancer screening.

Future work could focus on refining the segmentation techniques to improve the accuracy of archetype quantification further. Additionally, expanding the model to include other interpretable archetypes and validating it across a broader range of datasets could enhance

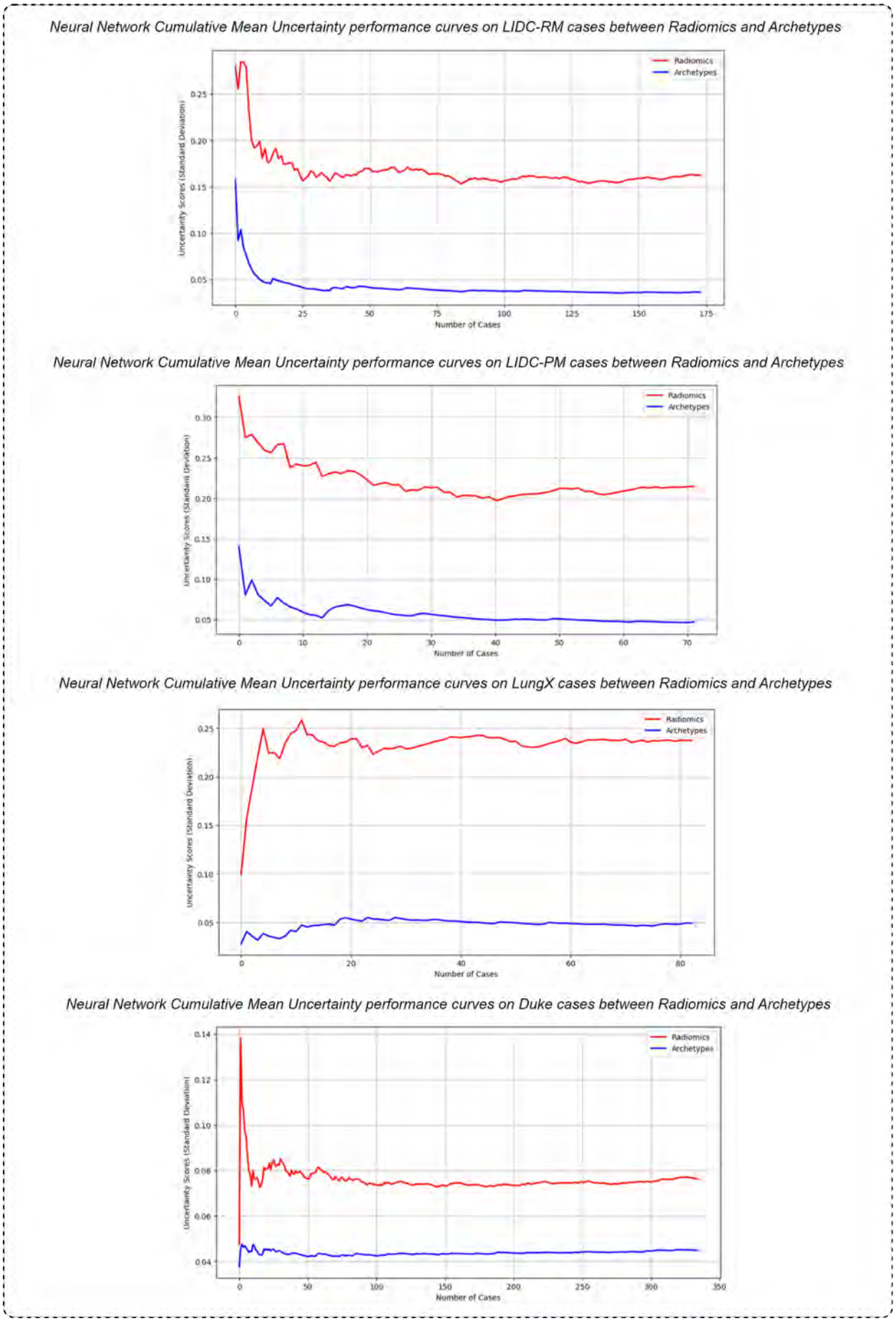



Figure 6: Neural Network Uncertainty Estimation Trends on different datasets

its generalizability and robustness. Integrating this approach with existing clinical workflows and assessing its impact on diagnostic decision-making and patient outcomes would be critical next steps in translating this research into practical clinical applications.

## 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

Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., et al., 2019. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature medicine 25, 954–961.

Armato III, S.G., Drukker, K., Li, F., Hadjiiski, L., Tourassi, G.D., Engelmann, R.M., Giger, M.L., Redmond, G., Farahani, K., Kirby, J.S., et al., 2016. Lungx challenge for computerized lung nodule classification. Journal of Medical Imaging 3, 044506–044506.

Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al., 2011. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics 38, 915–931.

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.

Bade, B.C., Cruz, C.S.D., 2020. Lung cancer 2020: epidemiology, etiology, and prevention. Clinics in chest medicine 41, 1–24.

Buty, M., Xu, Z., Gao, M., Bagci, U., Wu, A., Mollura, D.J., 2016. Characterization of lung nodule malignancy using hybrid shape and appearance features, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part I 19, Springer. pp. 662–670.

Cao, W., Wu, R., Cao, G., He, Z., 2020. A comprehensive review of computer-aided diagnosis of pulmonary nodules based on computed tomography scans. IEEE Access 8, 154007–154023.

Caspers, J., 2021. Translation of predictive modeling and ai into clinics: a question of trust. European Radiology 31, 4947–4948.

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K., 2019. This looks like that: deep learning for interpretable image recognition. Advances in neural information processing systems 32.

Choi, W., Oh, J.H., Riyahi, S., Liu, C.J., Jiang, F., Chen, W., White, C., Rimner, A., Mechalakos, J.G., Deasy, J.O., et al., 2018. Radiomics analysis of pulmonary nodules in low-dose ct for early detection of lung cancer. Medical physics 45, 1537–1549.

Choi, W., Xue, M., Lane, B.F., Kang, M.K., Patel, K., Regine, W.F., Klahr, P., Wang, J., Chen, S., D'Souza, W., et al., 2016. Individually optimized contrast-enhanced 4d-ct for radiotherapy simulation in pancreatic ductal adenocarcinoma. Medical physics 43, 5659–5666.

Ciompi, F., Jacobs, C., Scholten, E.T., van Riel, S.J., Wille, M.M., Prokop, M., van Ginneken, B., 2015. Automatic detection of spiculation of pulmonary nodules in computed tomography images, in: Medical Imaging 2015: Computer-Aided Diagnosis, SPIE. pp. 58–63.

Dhara, A.K., Mukhopadhyay, S., Saha, P., Garg, M., Khandelwal, N., 2016. Differential geometry-based techniques for characterization of boundary roughness of pulmonary nodules in ct images. International journal of computer assisted radiology and surgery 11, 337–349.

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.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: international conference on machine learning, PMLR. pp. 1050–1059.

Ghorbani, A., Wexler, J., Zou, J.Y., Kim, B., 2019. Towards automatic concept-based explanations. Advances in neural information processing systems 32.

Gould, M.K., Tang, T., Liu, I.L.A., Lee, J., Zheng, C., Danforth, K.N., Kosco, A.E., Di Fiore, J.L., Suh, D.E., 2015. Recent trends in the identification of incidental pulmonary nodules. American journal of respiratory and critical care medicine 192, 1208–1214.

Hawkins, S., Wang, H., Liu, Y., Garcia, A., Stringfield, O., Krewer, H., Li, Q., Cherezov, D., Gatenby, R.A., Balagurunathan, Y., et al., 2016. Predicting malignant nodules from screening ct scans. Journal of Thoracic Oncology 11, 2120–2128.

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H., Aerts, H.J., 2018. Artificial intelligence in radiology. Nature Reviews Cancer 18, 500–510.

Hou, Y., Zhang, Y.H., Bao, J., Bao, M.L., Yang, G., Shi, H.B., Song, Y., Zhang, Y.D., 2021. Artificial intelligence is a promising prospect for the detection of prostate cancer extracapsular extension with mpmri: a two-center comparative study. European Journal of Nuclear Medicine and Molecular Imaging 48, 3805–3816.

Kumar, D., Chung, A.G., Shaifee, M.J., Khalvati, F., Haider, M.A., Wong, A., 2017. Discovery radiomics for pathologically-proven computed tomography lung cancer prediction, in: Image Analysis and Recognition: 14th International Conference, ICIAR 2017, Montreal, QC, Canada, July 5–7, 2017, Proceedings 14, Springer. pp. 54–62.

Lambin, P., Leijenaar, R.T., Deist, T.M., Peerlings, J., De Jong, E.E., Van Timmeren, J., Sanduleanu, S., Larue, R.T., Even, A.J., Jochems, A., et al., 2017. Radiomics: the bridge between medical imaging and personalized medicine. Nature reviews Clinical oncology 14, 749–762.

Larici, A.R., Farchione, A., Franchi, P., Ciliberto, M., Cicchetti, G., Calandriello, L., Del Ciello, A., Bonomo, L., 2017. Lung nodules: size still matters. European respiratory review 26.

van Leeuwen, K.G., Schalekamp, S., Rutten, M.J., van Ginneken, B., de Rooij, M., 2021. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. European radiology 31, 3797–3804.

Liu, L., Dou, Q., Chen, H., Qin, J., Heng, P.A., 2019. Multi-task deep model with margin ranking loss for lung nodule analysis. IEEE transactions on medical imaging 39, 718–728.

Liu, Y., Balagurunathan, Y., Atwater, T., Antic, S., Li, Q., Walker, R.C., Smith, G.T., Massion, P.P., Schabath, M.B., Gillies, R.J., 2017. Radiological image traits predictive of cancer status in pulmonary nodules. Clinical Cancer Research 23, 1442–1449.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems 30.

MacMahon, H., Naidich, D.P., Goo, J.M., Lee, K.S., Leung, A.N., Mayo, J.R., Mehta, A.C., Ohno, Y., Powell, C.A., Prokop, M., et al., 2017. Guidelines for management of incidental pulmonary nodules detected on ct images: from the fleischner society 2017. Radiology 284, 228–243.

McKee, B.J., Regis, S.M., McKee, A.B., Flacke, S., Wald, C., 2016.

Performance of acr lung-rads in a clinical ct lung screening program. Journal of the American College of Radiology 13, R25–R29.

McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.S., Darzi, A., et al., 2020. International evaluation of an ai system for breast cancer screening. Nature 577, 89–94.

Niehaus, R., Stan Raicu, D., Furst, J., Armato, S., 2015. Toward understanding the size dependence of shape features for predicting spiculation in lung nodules for computer-aided diagnosis. Journal of digital imaging 28, 704–717.

Oh, Y., Park, S., Ye, J.C., 2020. Deep learning covid-19 features on cxr using limited training data sets. IEEE transactions on medical imaging 39, 2688–2700.

Reyes, M., Meier, R., Pereira, S., Silva, C.A., Dahlweid, F.M., Tengg-Kobligk, H.v., Summers, R.M., Wiest, R., 2020. On the interpretability of artificial intelligence in radiology: challenges and opportunities. Radiology: artificial intelligence 2, e190043.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144.

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.

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.

Shen, S., Han, S.X., Aberle, D.R., Bui, A.A., Hsu, W., 2019. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. Expert systems with applications 128, 84–95.

Swensen, S.J., Silverstein, M.D., Ilstrup, D.M., Schleck, C.D., Edell, E.S., 1997. The probability of malignancy in solitary pulmonary nodules: application to small radiologically indeterminate nodules. Archives of internal medicine 157, 849–855.

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.

Vezhnevets, V., Konouchine, V., 2005. Growcut: Interactive multi-label nd image segmentation by cellular automata, in: proc. of Graphicon, Citeseer. pp. 150–156.

Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE transactions on medical imaging 23, 903–921.

Xie, Y., Xia, Y., Zhang, J., Song, Y., Feng, D., Fulham, M., Cai, W., 2018. Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest ct. IEEE transactions on medical imaging 38, 991–1004.

Xie, Y., Zhang, J., Xia, Y., 2019. Semi-supervised adversarial model for benign–malignant lung nodule classification on chest ct. Medical image analysis 57, 237–248.

Yip, S.S., Parmar, C., Blezek, D., Estepar, R.S.J., Pieper, S., Kim, J., Aerts, H.J., 2017. Application of the 3d slicer chest imaging platform segmentation algorithm for large lung nodule delineation. PLoS One 12, e0178944.

Zhang, J., Xie, Y., Xia, Y., Shen, C., 2019. Attention residual learning for skin lesion classification. IEEE transactions on medical imaging 38, 2092–2103.

Zhao, Y., de Bock, G.H., Vliegenthart, R., van Klaveren, R.J., Wang, Y., Bogoni, L., de Jong, P.A., Mali, W.P., van Ooijen, P.M., Oudkerk, M., 2012. Performance of computer-aided detection of pulmonary nodules in low-dose ct: comparison with double reading by nodule volume. European radiology 22, 2076–2084.

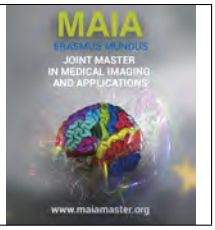
Zheng, X., Yao, Z., Huang, Y., Yu, Y., Wang, Y., Liu, Y., Mao, R., Li, F., Xiao, Y., Wang, Y., et al., 2020. Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer. Nature communications 11, 1236.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2921–2929.

Zwanenburg, A., Vallières, M., Abdalah, M.A., Aerts, H.J., Andrearczyk, V., Apte, A., Ashrafinia, S., Bakas, S., Beukinga, R.J., Boellaard, R., et al., 2020. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology 295, 328–338.

Medical Imaging and Applications

Master Thesis, June 2024




# Segmentation of Brain Cortex from Ultra-low-field MRI: A Prerequisite for Surface Reconstruction

Daniel Tweneboah Anyimadu[a,b], Dr. Emma C. Robinson[a]*, Dr. František Váša[b]*

*Joint Supervisors

[a]Department of Biomedical Engineering and Imaging Sciences (BMEIS), King's College London, UK
[b]Academic Neurosciences Centre, King's College London, UK

**Abstract**

Understanding the human cerebral cortex is vital due to its role in higher-level brain functions. However, its thin, convoluted structure makes imaging challenging, especially at ultra-low-field (ULF) magnetic resonance imaging (MRI) strengths. While artificial intelligence (AI) advancements have enhanced segmentation at high-field (HF) strengths, setting a baseline for cortical reconstruction; there is a paucity of research on cortex segmentation at ULF. This study adapts and fine-tunes AI-driven segmentation models including U-Net, SegResNet, and nnU-Net as well as investigates coupling these with advanced preprocessing techniques like resampling, bias-field correction, denoising, and skull-stripping for enhanced ULF cortical segmentation. The study utilized two comprehensive datasets, including adults (ages 20-69) and infants (ages 12-18 months), with both HF and ULF MRI to train and validate the models. Quantitative evaluations from this research demonstrated that nnU-Net outperformed other models, achieving the highest mean Dice Coefficient scores (0.9078 and 0.9132 in adults; 0.8472 in infants) and the lowest mean Hausdorff Distances (7.6419mm and 8.6077mm in adults; 9.7682mm in infants) across three ULF tested samples. Using the optimized BOUNTI (Brain vOlumetry and aUtomated parcellatioN for 3D feTal MRI) pipeline, adapted from the dHCP neonatal version, we successfully extracted cortical surfaces from ULF infant datasets, enabling detailed analysis of cortical surface area and morphology, which is vital for neurodevelopmental assessment. Our proposed nnU-Net segmentation framework therefore establishes a foundation for cortical surface reconstruction. It is accessible via *MScThesis*.

*Keywords:* *Brain Cortex*, *Ultra-low-field MRI*, *AI-driven Segmentation*, *Surface Reconstruction*.

## 1. Introduction

The cerebral cortex, is the brain's outer neural tissue layer in humans and mammals, playing a pivotal role in neural integration. It comprises four lobes: frontal, parietal, temporal, and occipital, collectively facilitating high-level brain functions such as sensory perception, motor control, and complex cognition (Javed et al., 2023; Molnár et al., 2019; Stiles and Jernigan, 2010). The brain's complexity renders direct examination time-consuming and skill-intensive, underscoring the need for medical imaging (Keith, 2023; Langen et al., 2017). Neuroimaging, has thus, become an essential procedure for exploring the structure or activity of the brain as well as for identifying changes linked to neurological disorders (Vasung et al., 2019).

Modern healthcare increasingly relies on magnetic resonance imaging (MRI) to noninvasively visualize structure and function of the brain, including the cerebral cortex. This technology has significantly enhanced our understanding of neural activities and brain development (Shetewi et al., 2020). MRI's safe, nonionizing, quantitative analysis capabilities, and 3D imaging make it ideal for advanced, AI-driven neurological diagnostics (Tocchio et al., 2015). However, the high costs (greater than £1 million) of high-field (1.5T or 3T) MRI scanners, limit their global accessibility, even in developed countries (Liu et al., 2021). For example, West Africa has only 84 MRI units for over 372 million people, with Ghana having 0.48 units per million and Nige-

ria 0.30 units per million (Ogbole et al., 2018).

Conversely, ultra-low-field (ULF) MRI scanners like the 0.064T Hyperfine Swoop promise a revolution in healthcare for low- and middle-income countries (Abate et al., 2024; Arnold et al., 2023). Affordable and portable, the Hyperfine operates from a standard electrical outlet, allowing bedside or field scans. Despite their potential in diagnosing conditions like strokes and tumours (Shoghli et al., 2023), these ULF MRI scanners grapple with issues such as low resolution, image noise, and extended scan times, which currently limit their broader use (Liu et al., 2021).

Artificial Intelligence (AI) (Holmes et al., 2004), on the other hand, is revolutionizing medical imaging, particularly in MRI technology, by improving image quality and reducing artifacts without the high costs associated with traditional HF MRI scanners (Ertl-Wagner and Wagner, 2023). Deep Learning (DL) Networks (Choi et al., 2020; Shen et al., 2017), are at the forefront of this transformation, using large datasets to refine and accelerate the development of ULF MRI, making advanced diagnostic capabilities more accessible and affordable (Islam et al., 2023).

DL-based convolutional neural networks (CNNs) (O'shea and Nash, 2015) and U-Nets (Ronneberger et al., 2015) are particularly beneficial for segmentation tasks due to their ability to learn complex patterns from large datasets. These models excel in differentiating between various brain tissues, such as cerebrospinal fluid (CSF), ventricles (VEN), white matter (WM), deep grey matter (grey matter subcortex; GMS), cortical grey matter (grey matter cortex; GMC), brainstem (BS), and cerebellum (CB). This capability enhances understanding of neurological conditions, enables precise mapping of functional areas, and facilitates detailed statistical analysis (Singh et al., 2022).

Historically, automatic segmentation algorithms have succeeded in identifying tissue types and anatomical features in MRI scans, relying heavily on clear tissue contrasts (Hamghalam et al., 2020; Jalab and Hasan, 2019). Key advancements include atlas-based segmentation (Cabezas et al., 2011), which utilizes pre-labeled atlases to guide the segmentation of new images; threshold methods (Al-Amri et al., 2010), which segment images by setting intensity value thresholds to separate regions; region-growing (Dehdasht-Heydari and Gholami, 2019), which expand regions based on similarity criteria from seed points; and watershed transformations (Kwon et al., 2016), which treat the image as a topographic surface and segment it by simulating water flow. Recently, CNNs like 3D Unet and DeepSCAN have outperformed traditional methods for lesion and brain segmentation in multiple sclerosis (McKinley et al., 2021). Additionally, machine learning super-resolution (SR) algorithms have improved segmentation accuracy and diagnostic performance in ULF MRI (Iglesias et al., 2022). Despite these advancements, there remains a paucity in research on applying these algorithms to ultra-low contrast images for cortical segmentation. Accurate segmentation of ULF cortex is therefore crucial as it provides essential brain maps for surface reconstruction.

Surface reconstruction in neuroimaging is a sophisticated computational process used to create a smooth, continuous, two-dimensional representation of the brain's surface from 3D volumetric MRI data (Ren et al., 2022). This technique is paramount in medical imaging as it facilitates a detailed examination of the brain's anatomy, offering relevant insights into its complex structures and functions. Specifically, for the cerebral cortex, the brain's outer layer, surface reconstruction allows clinicians and researchers to analyze cortical thickness, surface area, and gyrification patterns, which are vital for diagnosing and understanding a variety of neurological conditions, such as Alzheimer's disease, schizophrenia, and epilepsy (Fernández-Pena, 2023).

One of the most renowned techniques for cortical surface reconstruction is the "recon-all" pipeline from FreeSurfer (Fischl, 2012), a comprehensive suite used for the processing and analysis of brain MRI data. This pipeline involves a series of stages including motion correction, skull stripping, segmentation, normalization, and tessellation of the GMC, culminating in the creation of cortical surface models. While highly effective, the traditional "recon-all" pipeline is optimized for adult T1-weighted (T1w) images that provide high contrast and spatial resolution (Buxton et al., 1987),making it less effective with T2-weighted (T2w) or ULF MRI. This has spurred the creation of new surface reconstruction pipelines designed to overcome the challenges associated with lower-quality imaging modalities. Advanced segmentation models are central to these innovations, vital for accurate surface reconstruction, especially for precise differentiation between GMC and WM boundaries in ULF MRI.

This project addresses the challenges of ULF MRI by refining AI segmentation algorithms to enhance its capabilities. This approach broadens neuroimaging applications, lays the groundwork for surface reconstruction, and improves the diagnosis and understanding of neurological conditions, especially in populations where conventional HF MRI is less effective or accessible.

## 2. State of the art

Advancements in DL (Choi et al., 2020; Shen et al., 2017) **shown in Figure 1**, especially with CNNs (O'shea and Nash, 2015), U-Net (Ronneberger et al., 2015), and ResNet (He et al., 2016), have revolutionized medical image segmentation by significantly enhancing precision and reliability in partitioning images to identify and delineate anatomical structures. CNNs excel in medical imaging tasks due to their hierarchical feature learning, which captures both low-level and high-

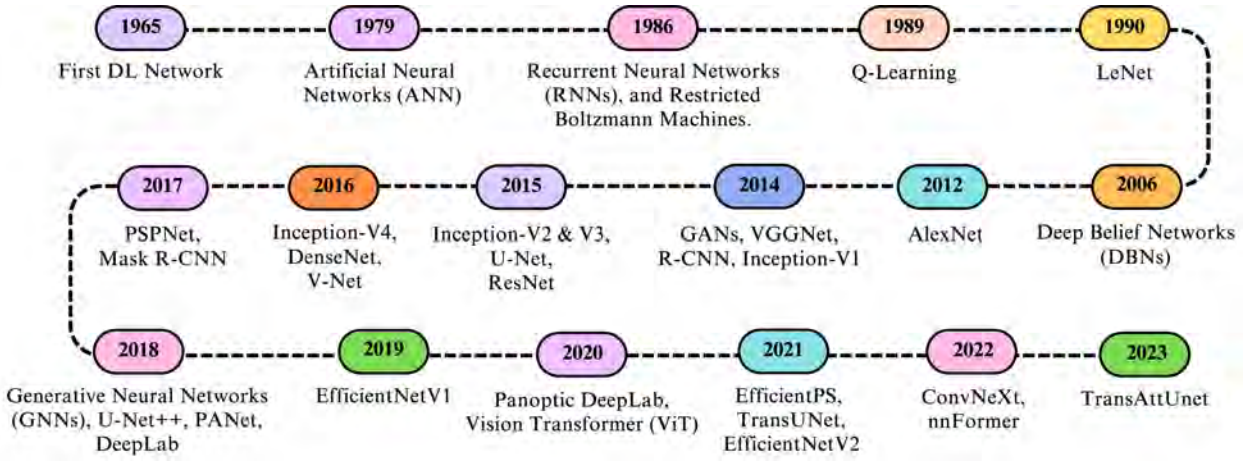



Figure 1: Timeline of the most popular DL algorithms in medical image segmentation (Rayed et al., 2024).

level features. U-Net's encoder-decoder structure with skip connections allows for efficient capture of spatial context and precise localization of anatomical features. Notable application include Lee et al. (2020), who used a Patch-wise U-net to segment healthy T1 OASIS and IBSR brain tissues, achieving an average Dice Similarity Coefficient (DSC) of 0.93 for CSF, GM, and WM in the OASIS dataset.

The ResNet architecture, developed by He et al. (2016), incorporates residual learning to improve the training of deep neural networks. It has been effectively used in conjunction with U-Net for enhancing segmentation tasks. For instance, Ramzan et al. (2020) used 3D CNNs with residual learning to segment T1w, registered T1w, and FLAIR images from several datasets, including MICCAI 2012, achieving DSCs of 0.879 for brain tissues segmented into 3 regions and 0.914 for 9 regions.

nnU-Net ("no-new-Net"), proposed by Isensee et al. (2021), is a self-adapting framework that dynamically configures its architecture based on dataset characteristics. It has set new standards in medical image segmentation by automating preprocessing, training, and post-processing steps. Baniasadi et al. (2023) employed CNNs from the nnU-Net framework to segment T1 images from multiple datasets, including the Human Connectome Project (HCP) young adults, obtaining an average DSC of 0.89 ± 0.04 for 30 brain regions.

These studies relied on ground-truth annotations, which are challenging to obtain due to their time-consuming nature and the requirement for expert knowledge (Jacob et al., 2021; Weese and Lorenz, 2016). SynthSeg (Billot et al., 2023), a CNN-based segmentation tool, helps to handle this constraint in neuroimaging, to achieve highly accurate segmentations across various brain MRI modalities, agnostic of contrast and resolution. The advent of advanced segmentation toolkits such as "segmentation-models-3D," (Solovyev et al., 2022) MONAI (Cardoso et al., 2022), nnU-Net (Isensee et al., 2021) and SynthSeg (Billot et al., 2023), has further improved segmentation performance.

Despite these advancements, there remains a significant gap in leveraging DL models for ULF brain cortical segmentation. Our project aims to segment cortical structures from ULF brain MRI scans by adapting and fine-tuning these models, coupled with advanced preprocessing techniques. This effort seeks to achieve improved segmentation outcomes, laying the foundation for further surface reconstruction.

### *2.1. Key Terminologies and Concepts Used*

In this project, **segmentation** involves partitioning medical images to identify anatomical structures. **Ultra-Low-Field (ULF) MRI** operates at lower magnetic field strengths (0.064T), offering affordability but with lower resolution challenges. **Patch-wise processing** divides images into smaller sections for detailed analysis. Key anatomical structures include cerebrospinal fluid **(CSF)**, ventricles **(VEN)**, white matter **(WM)**, grey matter subcortex **(GMS)**, grey matter cortex **(GMC)**, brainstem **(BS)**, and cerebellum **(CB)**. **Data transformation and augmentation** enhance training datasets by modifying images to improve model robustness. **nnU-Net** is a self-configuring deep learning framework for medical image segmentation, while **SegResNet** combines segmentation with residual learning. **Skip connections** help mitigate vanishing gradients, and **transposed convolutions** upsample feature maps. The **Adam optimizer and Dice loss function** are used for training, measuring overlap between predictions and true segmentations. **Dropout probability** prevents overfitting by deactivating neurons randomly. **Model checkpoints** save optimal model weights during training and **sliding window inference** aggregates predictions from overlapping patches. **Surface reconstruction** creates a smooth representation of the brain's surface from segmented MRI data for detailed anatomical analysis.

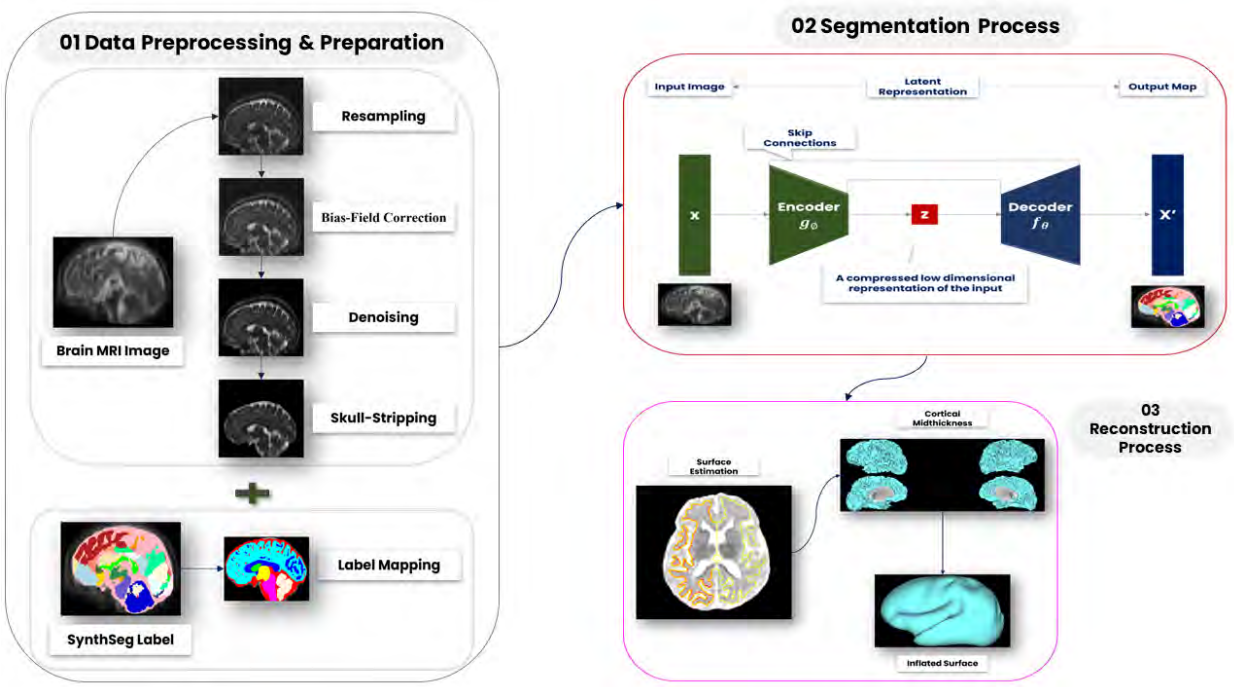



Figure 2: Overall project workflow including (1) data pre-processing, (2) segmentation, and (3) reconstruction processes.

## 3. Material and methods

Here, we first overview the MRI sequences, datasets used, and data preprocessing steps. Next, we explain the CNN/Unet architectures used in segmentation, detailing the selection, adaptation, and fine-tuning of - our selected models for ULF MRI scans and their evaluation methods. Finally, we elucidate the segmentation process and detail the reconstruction phase employing the surface reconstruction from the neonatal dHCP (developing Human Connectome Project) pipeline (Makropoulos et al., 2018). The project workflow is illustrated in **Figure 2**.

### *3.1. MRI Sequences / Image Dataset*

MRI sequences like T1w and T2w images are essential in neuroimaging. Both offer high anatomical detail and distinct contrasts due to different quantum relaxation processes. T1w is typically used for adult and infant segmentation, highlighting grey and white matter, while T2w is particularly useful for neonatal segmentation. These sequences are vital for effective image preprocessing and segmentation in both HF and ULF MRI. This study specifically utilizes the T2w MRI sequence **shown in Figure 3 (i)**.

The MRI scans used in this research come from a neurodevelopmental study that includes various scanner types and settings for thorough analysis. The dataset includes images from two primary sources: (1) HYPE Data - For this study, participants included 23 healthy adults, balanced by gender (2/3 male and 2/3 female), across five age-defined strata: 20-29, 30-39, 40-49, 50-59 and 60-69 years old. They were scanned using a HF GE Premier MRI scanner (3T, 1x1x1mm) and two identical Hyperfine Swoop ULF MRI scanners (64mT, 1.5x1.5x5mm) captured along three orthogonal planes (axial, sagittal, and coronal). The ULF scans were performed at the Centre for Neuroimaging Sciences (HFC) and Evelina Newborn Imaging Centre (HFE). Scanning times were about 45 minutes on the GE and 1 hour on the Hyperfine Swoop, covering both T1w and T2w scans. (2) Khula Data (Zieff et al., 2024)- We used a subset of the dataset containing ULF and HF T2w MRI scans of infants at 12 and 18 months, featuring 23 subjects for each scan type. The Khula Study, a multi-modal, multi-site longitudinal birth cohort, aims to characterize emerging executive functions in the first 1000 days of life in South Africa and Malawi. It enrolled 394 mothers from Gugulethu, Cape Town, and 507 mothers from Blantyre, providing valuable insights into early neurodevelopment. The Khula dataset offers a unique perspective on brain development in low- and middle-income countries, focusing on executive functions development and outcomes for children in these settings.

#### *3.1.1. Generation of Ground-Truth Annotations*

Ground-truth annotations are essential for evaluating the accuracy of segmentation models applied to brain MRI images. Due to the unavailability of manual labels, we used SynthSeg+ as described by Billot et al. (2023). SynthSeg+ generates precise segmentation la-

bels, providing highly accurate reference data for supervised learning. For each MRI scan, target labels were created using the "mri_synthseg" function in FreeSurfer (Fischl, 2012). These generated labels are vital for validating our segmentation algorithms. Integrating varied data sources and carefully preparing target labels, **depicted in Figure 3 (i)**, ensure that our tested algorithms perform well across diverse imaging scenarios, enhancing their applicability in clinical and research settings.

### *3.2. Dataset Preprocessing and Preparation*

#### *3.2.1. Image Resampling*

Before resampling, images were registered using FLIRT (FMRIB's Linear Image Registration Tool) (Smith et al., 2004), aligning ULF images to their HF counterparts. This process automatically resampled voxels. However, to maintain consistency across our dataset (MRI scans and their corresponding target labels), we used the "sitk.ResampleImageFilter()" function in SimpleITK toolkit (Yaniv et al., 2018). This method standardizes the voxel spacing to 1mm³ for each MRI scan, modifying original dimensions and spacings to preserve image proportion integrity through appropriate transformations. For target labels, we used nearest neighbor interpolation to avoid altering label values. Non-label images underwent B-spline interpolation to smooth pixel values. **Figure 3 (ii)** illustrates the sequence of brain MRI image preprocessing, starting with resampling as the initial step.

#### *3.2.2. Bias Field Correction*

To address spatial resolution variability and inconsistent image intensities in the HYPE and Khula datasets, a streamlined bias field correction (BFC) and denoising workflow was employed. Using the "n4_bias_field_correction" algorithm from the ANTsPyNet library (Tustison et al., 2010, 2021), we initially masked each image to create a binary image, segmenting the brain tissue from the background. The BFC refines images in three steps: unblurring brightness levels, adjusting the Gaussian brightness distribution, and smoothing adjustments with B-spline modeling. This process effectively corrects the bias field, enhancing the uniformity and clarity of the scans **Figure 3 (ii)** depicts BFC as the second stage.

#### *3.2.3. Denoising*

For the denoising phase, the ANTsPyNet library's functionality "ants.denoise_image" was used post-BFC. This procedure employs the same mask utilized in the bias correction step to ensure specificity and effectiveness. This streamlined procedure not only improved the homogeneity and clarity of the images but also optimized them for the detailed segmentation tasks that followed, relevant for accurate and reliable neuroimaging analysis. Denoising is highlighted as the third stage in **Figure 3 (ii)**.

#### *3.2.4. Skull Stripping*

Skull stripping is an essential preprocessing step in MRI analysis to remove non-brain tissues and isolate the brain matter. We used the SynthStrip tool (Hoopes et al., 2022) from FreeSurfer for this task. SynthStrip, utilizing a DL approach, effectively handles images from various modalities and resolutions. The "mri_synthstrip" command was executed to remove the skull and save the stripped image. This efficient process preserves the quality of neural structures, essential for accurate subsequent analysis. **Figure 3 (ii)** depicts how images are skull-stripped in the last phase of the preprocessing.

#### *3.2.5. Label Mapping*

As a vital data preparatory step, we merged the 99 unique regions identified by SynthSeg+ into seven categories: CSF, VEN, WM, GMS, GMC, BS, and CB. This streamlined approach simplifies the raw MRI data, making it more manageable and enhancing the AI models' ability to focus on cortical structure segmentation, central to our project's goals.

Label mapping details include: (1) CSF: encompasses the 3rd ventricle, 4th ventricle, and surrounding CSF; (2) VEN: includes the left and right lateral ventricles, along with their inferior lateral aspects; (3) WM: covers the left and right cerebral white matter; (4) GMS: consists of key subcortical structures like the thalamus, caudate, putamen, pallidum, hippocampus, amygdala, accumbens area, and ventral diencephalon on both sides; (5) GMC: includes left and right cortex, including all cortical regions specified in the data; (6) BS: focused on the entire brain stem; (7) CB: separates both the left and right cerebellum white matter and cortex. **Figure 3 (iii)** depicts the conversion of ground-truth annotations from SynthSeg+ into seven distinct tissue classes.

#### *3.2.6. Data Splitting and Set-up*

Our pre-processed datasets were divided into training, validation, and testing sets, each containing 16, 4, and 3 samples respectively, drawn from both the HYPE and Khula datasets. These datasets include paired HF (3T) and ULF (64mT) MRI data, with the HYPE dataset comprising 1 set of HF data and 2 sets of ULF data (HFC and HFE), and the Khula dataset containing 1 set of HF data and 1 set of ULF data. Given the risk of model overfitting, particularly with small sample sizes, we validated our models using the unseen test set and employed patch-based training techniques on most of our models to mitigate overfitting in small datasets. This approach not only facilitates effective validation but also diversifies the data provided to the model during training, reducing the risk of overfitting.

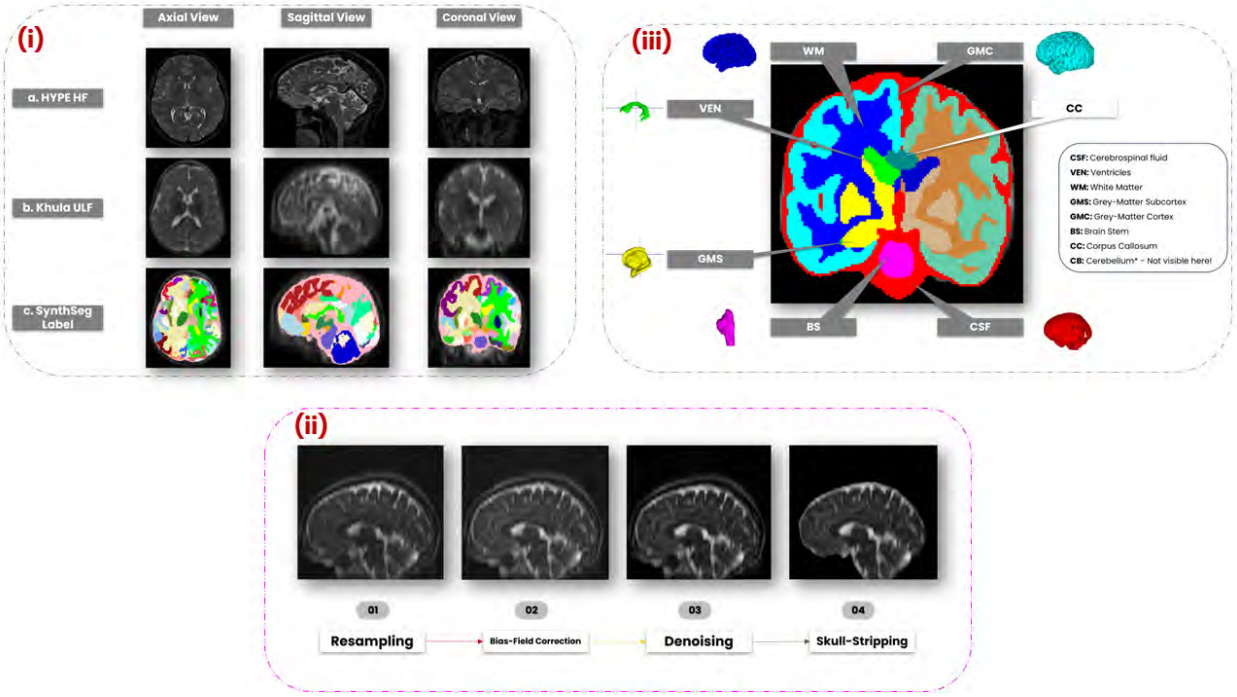



Figure 3: (i) Sample-T2w HF and ULF MRI images with SynthSeg-generated labels (for Khula ULF). (ii) Sample HYPE-HF Brain MRI image pre-processing workflow. (iii) Tissue class label mapping.

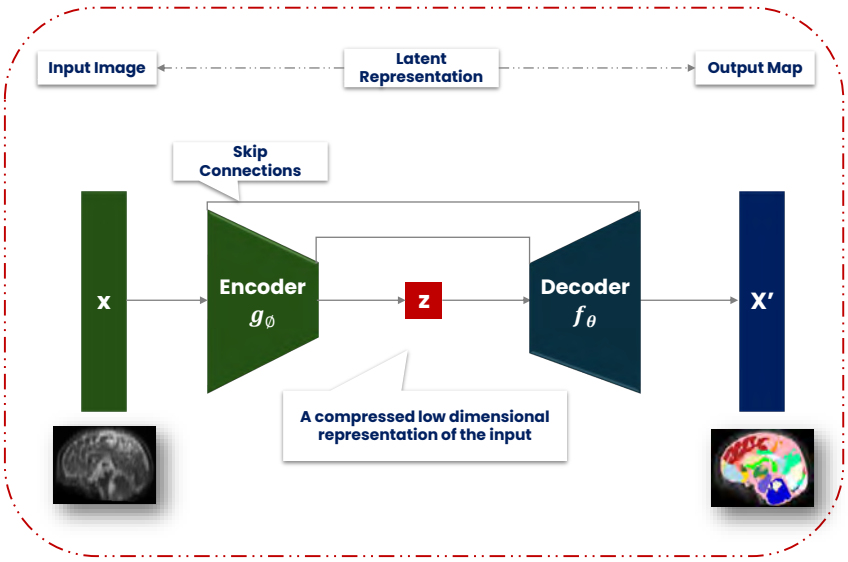



Figure 4: A simplified diagram of U-Net architecture.

### *3.3. CNN/Unet Architecture Overview*

CNNs, comprising convolutional layers, pooling, and nonlinear activations, form the backbone of many AI-driven image segmentation models, including the widely used U-Net architecture. **Figure 4** depicts a U-Net architecture which uses skip connections (to concatenate low-level and high-level features) between symmetric encoder and decoders to preserve spatial information. Encoder($z = g_\emptyset(x)$): compresses input $x$ into a latent-space representation $z$. Decoder($x' = f_\theta(z)$): predicts the output $x'$ from $z$, using transposed convolutions (which reconstruct the original image size from

a latent-space representation) to create a detailed segmentation map. These features influenced our choice of segmentation models based on U-Net architectures.

### *3.4. Segmentation Models and Model Training*

With detailed descriptions of the selected models provided in the following sections, we adapted and fine-tuned several U-Net based segmentation models, incorporating data augmentation, network adjustments, training epochs, and learning rate modifications. We began with a baseline U-Net Segmentation Model with a ResNet34 (He et al., 2016) backbone for training. Next, we explored MONAI-based toolkits (Cardoso et al., 2022), which offer 3D-UNet and SegResNet architectures. SegResNet, unlike traditional U-Net, uses ResNet-inspired residual connections for information propagation and mitigating the vanishing gradient problem. It also introduces skip-connections across resolutions, improving multiscale feature capture. These architectures are complemented by robust data transformation and augmentation utilities within MONAI, facilitating effective data preparation and augmentation for improved model generalization and performance. Lastly, we integrated the nnU-Net toolkit (Isensee et al., 2021, 2019) for our task, which provides a self-adapting framework, reducing manual setup and efficiently adapting to diverse imaging conditions. To monitor and fine-tune our models' performance during training, we used the Dice Score Coefficient (DSC) (Bertels et al., 2019) to assess the spatial overlap between predicted and ground truth segmentations.

#### *3.4.1. 3D Segmentation Model with ResNet34 Backbone*

The "segmentation-models-3D" (SM3D) repository (Solovyev et al., 2022) tailored for Keras (Jin et al., 2019), TensorFlow (Abadi et al., 2016) and PyTorch (Paszke et al., 2019) was used to implement our 3D U-Net model with a ResNet34 backbone. The original 2D U-Net architecture (depicted in **Figure 5**) which was adapted in the SM3D toolkit for 3D volumetric data, comprises a contracting path (left side) and an expansive path (right side). The contracting path captures context through repeated applications of two 3x3 convolutions, each followed by a rectified linear unit (ReLU) and a 2x2 max pooling operation with stride 2 for downsampling, doubling the number of feature channels at each step. The expansive path ensures precise localization via upsampling, 2x2 convolutions, concatenation with cropped feature maps from the contracting path, and two 3x3 convolutions, each followed by a ReLU. A final 1x1 convolution maps each 64-component feature vector to the desired number of classes. The network includes 23 convolutional layers in total.

We employed a 3D U-Net model with a ResNet34 backbone from the SM3D toolkit for brain tissue segmentation on ULF MRI scans. Our implementation comprised several key steps to ensure optimal performance. Firstly, we resized data dimensions to match the model's requirements, given its preference for dimensions divisible by 32 due to its lowest resolution of 32x32x32 pixels. Following this, we performed data preprocessing, reshaping image dimensions to accommodate a single channel for the input brain MRI image and 8 channels for the target label, representing the segmentation output map. Data scaling was conducted using the min-max function to standardize it and prevent bias during model training. Additionally, we prepared training and validation splits to effectively monitor model performance during epoch-wise training. The segmentation model was then set up, and subsequently, model training commenced. We configured the model with the Adam optimizer (Zhang, 2018) and Dice loss function to optimize accuracy and Dice Similarity Coefficient (DSC) metrics. The Adam optimizer's adaptive learning rate adjustment facilitated efficient training, while the Dice loss function measured the overlap between predicted and ground truth segmentation masks, ensuring precise segmentation. Finally, we implemented callbacks for model checkpoints to save the best weights during training.

#### *3.4.2. MONAI-based Segmentation Architectures*

The MONAI framework (Cardoso et al., 2022), a PyTorch-based platform, facilitates the development and deployment of medical AI models, offering architectures like U-Net and SegResNet tailored for tasks such as image segmentation. With its extensive suite of data transformations and utilities, MONAI streamlines preprocessing, training, and validation of medical images, ensuring reproducibility and deterministic operations. Leveraging cropped patches during training and employing sliding window inference, we adopt a patch-wise segmentation approach. This method divides input images into overlapping patches, segments each patch individually, and aggregates results to generate the final segmentation map, enhancing accuracy and efficiency in healthcare AI applications.

(a) U-Net

Utilizing the U-Net architecture from the MONAI framework, our implementation includes essential preprocessing and data transformations to enhance the quality and responsiveness of our data. The 3D U-Net architecture, with 1 input channel and 8 output channels (for each of the eight tissue classes), progressively doubles channel depths from 16 to 256, incorporating batch normalization for efficient learning. The model is fine-tuned to handle the constraints of ULF MRI images by adjusting the learning rate, increasing training epochs, and modifying layer depth for better feature extraction based on initial training performance on HF images. Our workflow for HF and ULF data involves setting up a deterministic environment for reproducibility,

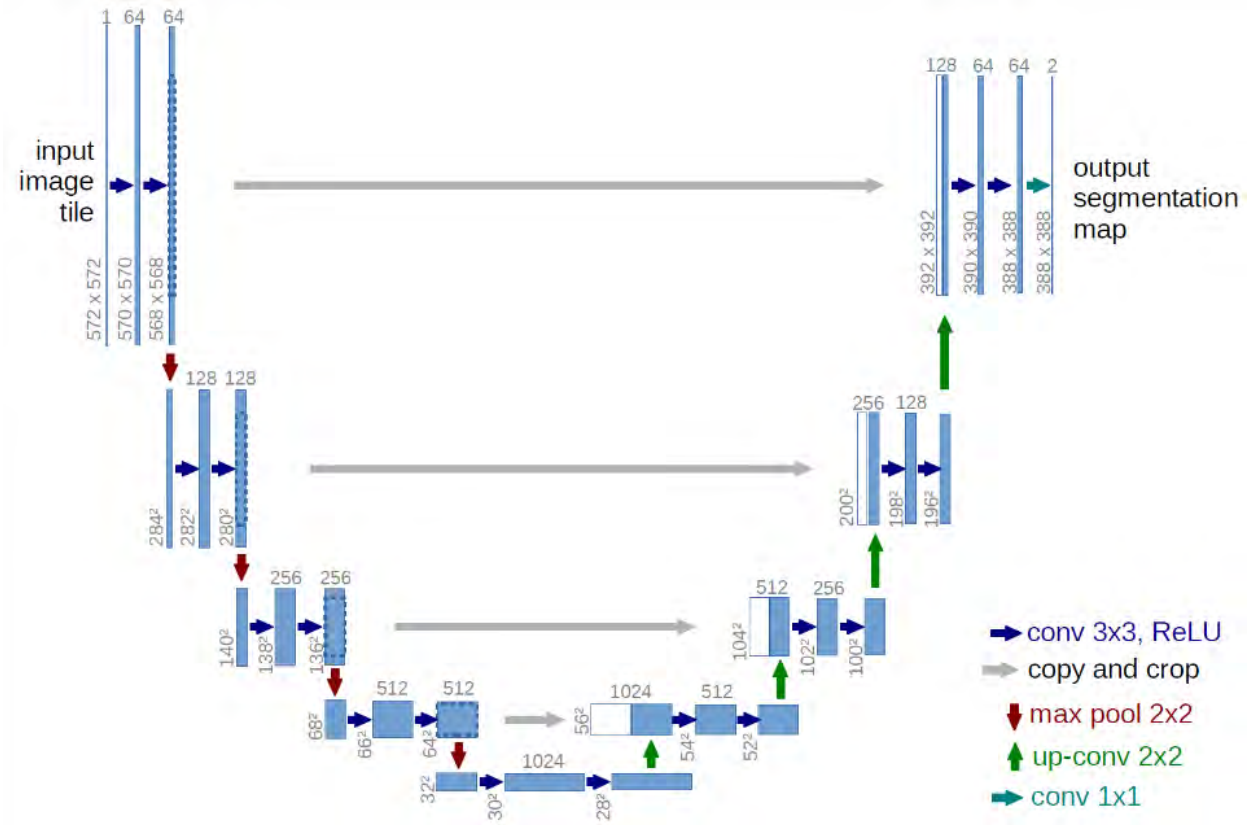



Figure 5: 2D U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations (Ronneberger et al., 2015).

comprehensive data augmentations (normalization, orientation correction, random crops), and robust training protocols using Adam optimization and Dice loss. We evaluate model performance using the DSC through a sliding window inference method, ensuring high fidelity in segmenting brain structures.

(b) SegResNet

Similarly, we utilized the SegResNet architecture from the MONAI framework to segment brain MRI images. The 3D SegResNet architecture, with 1 input channel, 8 output channels, initial filters set at 16, a dropout probability of 20% (by randomly deactivating neurons during training to improve model generalizability), and configured blocks for downsampling (1, 2, 2, 4) and upsampling (1, 1, 1), efficiently processes cropped MRI images. To adapt to ULF MRI constraints, we fine-tuned the SegResNet model by adjusting filter sizes and layer configurations based on initial performance on HF training data. This iterative refinement optimizes the model for lower resolution and contrast of ULF MRI scans. Key steps for our implementation included, setting up a deterministic environment for reproducibility, comprehensive data handling (loading, channel reordering, type conversion, orientation setting, uniform spacing), and augmentations (random cropping, flipping, intensity normalization, scaling, shifting). Robust training protocols using Adam optimization and Dice loss were implemented. Model performance is evaluated using DSC through a sliding window inference method.

*3.4.3. nnU-Net*

nnU-Net (Isensee et al., 2021, 2019) is a self-adapting segmentation framework that automates setup for medical image segmentation. It dynamically configures architecture based on dataset specifics using a dataset fingerprint, optimizing the use of 2D and 3D convolutions, depth, and filter number, while also training data patch-wise via a five-fold cross-validation. The framework standardizes data preprocessing by choosing normalization and resampling methods and employs extensive data augmentation (rotations, scaling, elastic deformations, gamma adjustments) to enhance robustness. Training uses a combined Dice and cross-entropy loss function with stochastic gradient descent (SGD) and adapts learning rates and early stopping to improve performance and prevent overfitting. During inference, nnU-Net employs a sliding window approach and may use ensemble methods either via hard voting (choosing the class with the highest probability for each pixel) or soft voting (averaging probabilities and selecting the class with the maximum probability) to refine segmen-

tation accuracy, incorporating post-processing steps for optimal results.

Employing nnU-Net in our project involved: (1) Data Preparation and Preprocessing: ensured data integrity and compatibility with nnU-Net using structured handling, Z-score normalization, and resampling to match median image size and spacing. (2) Model Training and Evaluation: initially patch-wise trained on HF images to establish a baseline, followed by fine-tuning for ULF MRI datasets with increased augmentation. Both 2D and 3D models were trained, and the best model was selected through an ensemble approach for inference. (3) Performance Evaluation: assessed using DSC to ensure high precision in tissue segmentation. The U-Net architecture used here handles both 2D and 3D inputs. The 2D U-Net begins with 32-filter convolutions, increasing to 512, and ends with a 1x1 convolutional output layer, trained with a batch size of 156. The 3D U-Net starts with 32-filter convolutions, increasing to 320, and ends with a 1x1x1 convolutional output layer with softmax activation, trained with a batch size of 2. Predictions are generated by forming an ensemble of learned features from both dimensions.

### *3.5. Evaluation Metrics*

To quantitatively evaluate the performance of our segmentation models on the test data, we utilized several metrics: DSC, Hausdorff Distance (HD) (Huttenlocher et al., 1993), and Average Volumetric Difference (AVD). **Figure 6** provides a visual representation of the metrics.

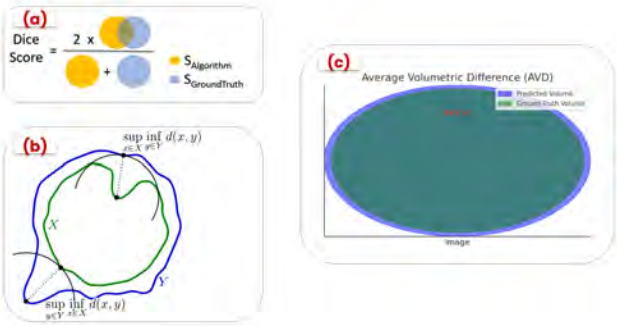



Figure 6: (a) shows the DSC overlap between predicted (yellow) and ground-truth (blue) segmentations, (b) illustrates the HD between ground-truth (X; green) and predicted (Y; blue) segmentations, and (c) depicts the AVD between predicted (blue) and ground-truth (green) segmentations.

(i) Dice Coefficient (DSC): Measures spatial overlap between the predicted and ground truth segmentations. A higher DSC indicates better accuracy, with values ranging from 0 to 1, where 1 denotes perfect overlap.

$$\mathrm{DSC} = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{1}$$

Where A: The set of predicted voxels. B: The set of ground-truth voxels. $|A \cap B|$ The number of voxels that are common between the predicted segmentation and the ground-truth. $|A| + |B|$: The total number of voxels in both the predicted segmentation and the ground truth.

(ii) Hausdorff Distance (HD): Measures the maximum distance between the boundaries of the predicted and ground-truth segmentations. A lower HD indicates more similar geometry. It is measured in units of distance (millimeters; mm) and ranges from 0 (perfect boundary match) to higher values (greater boundary discrepancy).

$$HD(A, B) = \max\left\{\sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(b, a)\right\} \tag{2}$$

Where A: The set of boundary points of the predicted segmentation. B: The set of boundary points of the ground-truth segmentation. d(a, b): The Euclidean distance between points a and b. sup: The supremum (or least upper bound). inf: The infimum (or greatest lower bound).

(iii) Average Volumetric Difference (AVD): Calculates the average absolute volume difference between predicted and ground-truth segmentations. A lower AVD indicates better volumetric agreement. It is measured in units of volume and ranges from 0 (perfect agreement) to higher values (greater discrepancy).

$$\mathrm{AVD} = \frac{1}{N} \sum_{i=1}^{N} |V_{\mathrm{pred},i} - V_{\mathrm{gt},i}| \tag{3}$$

N: The number of segmented regions or classes. Vpred,i: The volume of the predicted segmentation for region i. Vgt,i: The volume of the ground-truth segmentation for region i.

### *3.6. Segmentation Process: Conducted Experiments*

We conducted two experiments: (1) HF MRI Training and (2) ULF MRI Training, using the selected architectures (Table 1): 3D-Unet with ResNet34 backbone, MONAI architectures (U-Net, SegResNet), and nnU-Net. The experiments segmented pre-processed brain volumetric MRI data into tissue classes (CSF, VEN, WM, GMS, GMC, BS, CB) based on aggregated SynthSeg+ labels in a supervised DL approach. The models were trained extensively using 16 samples and validated with 4 samples over 150 to 250 epochs to optimize accuracy and adaptability to different field strengths. During training, performance was monitored using pre-defined evaluation metrics (DSC, HD, AVD) to guide parameter adjustments. Initially training on HF data assessed model performance and informed adjustments for optimal results on ULF MRI. After training, the models predicted segmentation on 3 unseen data samples. These predictions were quantitatively assessed against SynthSeg+ labels using DSC, HD, and AVD. Additionally,

qualitative quality control (QC) was performed by visualising images with ITK-SNAP (Yushkevich et al., 2016).

Table 1: Comparison of Segmentation Architectures

| Architecture | Filters | Activation |
|---|---|---|
| SM3D | 64, 128, 256, 512, 1024 | Softmax |
| MONAI 3DUNet | 32, 64, 128, 256, 320 | PReLU |
| MONAI SegResNet* | 16 | PReLU |
| nnU-Net 2D | 32, 64, 128, 256, 512 | LeakyReLU |
| nnU-Net 3D | 32, 64, 128, 256, 320 | LeakyReLU |

*Using initial filters of 16, the model downsamples with blocks [1, 2, 2, 4] and upsamples with blocks [1, 1, 1].

### *3.7. Cortical Surface Reconstruction with dHCP Pipeline*

Building on the robust segmentation results, we adopted the approach outlined by Makropoulos et al. (2018) for cortical surface reconstruction. This method, developed as part of the dHCP, involves fitting surfaces to our segmentations using the neonatal surface code by Schuh et al. (2017). To achieve this, we refined the cortical surface extraction process using a method adapted from the dHCP neonatal pipeline developed by Uus et al. (2023). This is accessible publicly at(`https://hub.docker.com/r/fetalsvrtk/segmentation`, tag brain_bounti_tissue). Our optimized version, (fetalsvrtk/surface:daniel_v1), was then employed to precisely extract cortical surfaces **demonstrated in Figure 9**. For the surface fitting process to function optimally, a corpus callosum mask was essential. Consequently, we manually annotated the corpus callosum in two samples using ITK-SNAP as a proof-of-concept. This integrated approach ensures accurate and reliable extraction of cortical surfaces **demonstrated in Figure 10**.

## 4. Results

### *4.1. Quantitative/Qualitative Analysis*

Here, highlighted results are for easy glance. For HF test results, mean model performances are in Tables 2 and 3, and tissue-specific results are in Tables 7 and 8. For ULF test results, mean performances are in Tables 4-6, and tissue-specific results are in Tables 9-11. Figures 7, 11, 12, 13 and 14 illustrate visual representations. Figure 8 provides a qualitative analysis of the segmentation models on Khula-ULF.

Table 2: Models Performance on HYPE-HF (Mean DSC, Mean HD, Mean AVD)

| MODEL | DSC | HD | AVD |
|---|---|---|---|
| nnU-Net | **0.9356** | **11.1188** | **0.0138** |
| MONAI 3DUNet | 0.9241 | 14.9077 | 0.0165 |
| MONAI SegResNet | 0.8959 | 23.2476 | 0.0393 |
| Unet(ResNet34) | 0.9019 | 34.6234 | 0.0387 |

Table 3: Models Performance on Khula-HF (Mean DSC, Mean HD, Mean AVD)

| MODEL | DSC | HD | AVD |
|---|---|---|---|
| nnU-Net | **0.9103** | **11.4457** | 0.0297 |
| MONAI 3DUNet | 0.8946 | 13.3780 | **0.0266** |
| MONAI SegResNet | 0.8586 | 47.9888 | 0.0949 |

Table 4: Models Performance on HYPE-ULF (HFC Samples)

| MODEL | DSC | HD | AVD |
|---|---|---|---|
| nnU-Net | **0.9078** | **7.6419** | **0.0357** |
| MONAI 3DUNet | 0.9047 | 9.8645 | 0.0377 |
| MONAI SegResNet | 0.8779 | 16.0166 | 0.0510 |

Table 5: Models Performance on HYPE-ULF (HFE Samples)

| MODEL | DSC | HD | AVD |
|---|---|---|---|
| nnU-Net | **0.9132** | **8.6077** | 0.0302 |
| MONAI 3DUNet | 0.9067 | 12.1760 | **0.0258** |
| MONAI SegResNet | 0.8857 | 11.9314 | 0.0573 |

Table 6: Models Performance on Khula-ULF (Mean DSC, Mean HD, Mean AVD)

| MODEL | DSC | HD | AVD |
|---|---|---|---|
| nnU-Net | **0.8472** | **9.7682** | 0.1180 |
| MONAI 3DUNet | 0.8432 | 15.2485 | **0.1033** |
| MONAI SegResNet | 0.8200 | 10.4125 | 0.1359 |

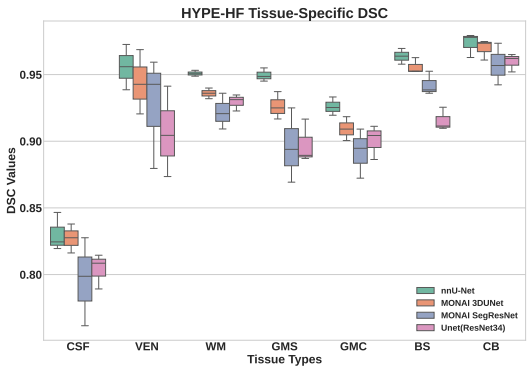



Figure 7: Illustration of DSC values for different brain tissues predicted by various models. Box plots show individual subject variations.

Table 7: Tissue-Specific DSC Performance of Models on HYPE-HF

| MODEL | SUBJECT | CSF | VEN | WM | GMS | GMC | BS | CB |
|---|---|---|---|---|---|---|---|---|
| nuU-Net | 1 | 0.8195 | 0.9559 | 0.9506 | 0.9486 | **0.9253** | 0.9638 | 0.9779 |
| | 2 | 0.8466 | 0.9762 | 0.9488 | 0.9451 | **0.9195** | 0.9578 | 0.9628 |
| | 3 | 0.8245 | 0.9386 | 0.9532 | 0.9550 | **0.9332** | 0.9696 | 0.9793 |
| MONAI 3DUNet | 1 | 0.8276 | 0.9427 | 0.9360 | 0.9250 | 0.9091 | 0.9525 | 0.9735 |
| | 2 | 0.8379 | 0.9687 | 0.9319 | 0.9167 | 0.9005 | 0.9524 | 0.9609 |
| | 3 | 0.8162 | 0.9205 | 0.9399 | 0.9372 | 0.9184 | 0.9627 | 0.9748 |
| MONAI SegResNet | 1 | 0.8276 | 0.9427 | 0.9360 | 0.9250 | 0.9091 | 0.9525 | 0.9735 |
| | 2 | 0.7987 | 0.9593 | 0.9092 | 0.8693 | 0.8723 | 0.9360 | 0.9423 |
| | 3 | 0.7616 | 0.8796 | 0.9207 | 0.8939 | 0.8947 | 0.9385 | 0.9568 |
| Unet(ResNet34) | 1 | 0.8085 | 0.9044 | 0.9312 | 0.8894 | 0.9043 | 0.9114 | 0.9621 |
| | 2 | 0.8145 | 0.9413 | 0.9227 | 0.8871 | 0.8863 | 0.9097 | 0.9520 |
| | 3 | 0.7892 | 0.8735 | 0.9347 | 0.9166 | 0.9112 | 0.9255 | 0.9650 |

Table 8: Tissue-Specific DSC Performance of Models on Khula-HF

| MODEL | SUBJECT | CSF | VEN | WM | GMS | GMC | BS | CB |
|---|---|---|---|---|---|---|---|---|
| nuU-Net | 1 | 0.7841 | 0.9559 | 0.9251 | 0.9469 | **0.9204** | 0.9506 | 0.9826 |
| | 2 | 0.7746 | 0.9207 | 0.9199 | 0.9483 | **0.9141** | 0.9546 | 0.9812 |
| | 3 | 0.7617 | 0.9374 | 0.9374 | 0.9100 | **0.9007** | 0.8590 | 0.9686 |
| MONAI 3DUNet | 1 | 0.7439 | 0.9505 | 0.9022 | 0.9345 | 0.8949 | 0.9346 | 0.9741 |
| | 2 | 0.7364 | 0.9105 | 0.8964 | 0.9336 | 0.8883 | 0.9490 | 0.9755 |
| | 3 | 0.7368 | 0.9306 | 0.8792 | 0.9043 | 0.8731 | 0.8729 | 0.9650 |
| MONAI SegResNet | 1 | 0.6751 | 0.9317 | 0.8757 | 0.8960 | 0.8630 | 0.9363 | 0.9592 |
| | 2 | 0.6607 | 0.8622 | 0.8529 | 0.8852 | 0.8446 | 0.9273 | 0.9493 |
| | 3 | 0.6758 | 0.9060 | 0.8375 | 0.8508 | 0.8362 | 0.8495 | 0.9563 |

Table 9: Tissue-Specific DSC Performance of Models on HYPE-ULF (HFC Samples)

| MODEL | SUBJECT | CSF | VEN | WM | GMS | GMC | BS | CB |
|---|---|---|---|---|---|---|---|---|
| nuU-Net | 1 | 0.7823 | 0.7823 | 0.7823 | 0.9269 | **0.8996** | 0.9532 | 0.9668 |
| | 2 | 0.7780 | 0.7780 | 0.7780 | 0.7780 | **0.8977** | 0.9599 | 0.9613 |
| | 3 | 0.7358 | 0.8986 | 0.9323 | 0.9138 | **0.8867** | 0.9494 | 0.9552 |
| MONAI 3DUNet | 1 | 0.8122 | 0.9335 | 0.9205 | 0.9024 | 0.8790 | 0.9554 | 0.9531 |
| | 2 | 0.8255 | 0.9514 | 0.9162 | 0.8910 | 0.8776 | 0.9497 | 0.9488 |
| | 3 | 0.7874 | 0.8958 | 0.9191 | 0.9014 | 0.8774 | 0.9461 | 0.9564 |
| MONAI SegResNet | 1 | 0.7628 | 0.9053 | 0.9028 | 0.8718 | 0.8534 | 0.9413 | 0.9542 |
| | 2 | 0.7756 | 0.9332 | 0.8949 | 0.8576 | 0.8427 | 0.9352 | 0.9371 |
| | 3 | 0.7416 | 0.8512 | 0.9005 | 0.8729 | 0.8478 | 0.9271 | 0.9266 |

Table 10: Tissue-Specific DSC Performance of Models on HYPE-ULF (HFE Samples)

| MODEL | SUBJECT | CSF | VEN | WM | GMS | GMC | BS | CB |
|---|---|---|---|---|---|---|---|---|
| nuU-Net | 1 | 0.7847 | 0.9325 | 0.9378 | 0.9211 | **0.8986** | 0.9526 | 0.9663 |
| | 2 | 0.8110 | 0.9486 | 0.9388 | 0.9189 | **0.9042** | 0.9500 | 0.9624 |
| | 3 | 0.7374 | 0.9187 | 0.9364 | 0.9366 | **0.8942** | 0.9616 | 0.9609 |
| MONAI 3DUNet | 1 | 0.8157 | 0.9196 | 0.9201 | 0.9076 | 0.8815 | 0.9491 | 0.9587 |
| | 2 | 0.8302 | 0.9524 | 0.9178 | 0.9063 | 0.8796 | 0.9408 | 0.9497 |
| | 3 | 0.7812 | 0.9108 | 0.9190 | 0.9247 | 0.8764 | 0.9502 | 0.9503 |
| MONAI SegResNet | 1 | 0.7793 | 0.9045 | 0.9045 | 0.8781 | 0.8579 | 0.9435 | 0.9545 |
| | 2 | 0.7952 | 0.9426 | 0.9018 | 0.8677 | 0.8584 | 0.9410 | 0.9460 |
| | 3 | 0.7540 | 0.8520 | 0.8992 | 0.8742 | 0.8511 | 0.9478 | 0.9441 |

Table 11: Tissue-Specific DSC Performance of Models on Khula-ULF

| MODEL | SUBJECT | CSF | VEN | WM | GMS | GMC | BS | CB |
|---|---|---|---|---|---|---|---|---|
| nuU-Net | 1 | 0.7599 | 0.9526 | 0.9141 | 0.9241 | **0.8928** | 0.9504 | 0.9481 |
| | 2 | 0.6165 | 0.8768 | 0.8797 | 0.8973 | **0.8474** | 0.7143 | 0.8990 |
| | 3 | 0.6556 | 0.9005 | 0.8801 | 0.8818 | **0.8513** | 0.7025 | 0.8473 |
| MONAI 3DUNet | 1 | 0.7311 | 0.9476 | 0.8831 | 0.9076 | 0.8593 | 0.9394 | 0.9470 |
| | 2 | 0.6264 | 0.8591 | 0.8572 | 0.8996 | 0.8188 | 0.9101 | 0.9275 |
| | 3 | 0.6276 | 0.9002 | 0.8486 | 0.8619 | 0.8226 | 0.6937 | 0.8384 |
| MONAI SegResNet | 1 | 0.6708 | 0.9318 | 0.8623 | 0.8810 | 0.8357 | 0.9251 | 0.9306 |
| | 2 | 0.5821 | 0.8473 | 0.8250 | 0.8643 | 0.7950 | 0.8809 | 0.9119 |
| | 3 | 0.5894 | 0.8810 | 0.8276 | 0.8368 | 0.8057 | 0.7015 | 0.8341 |

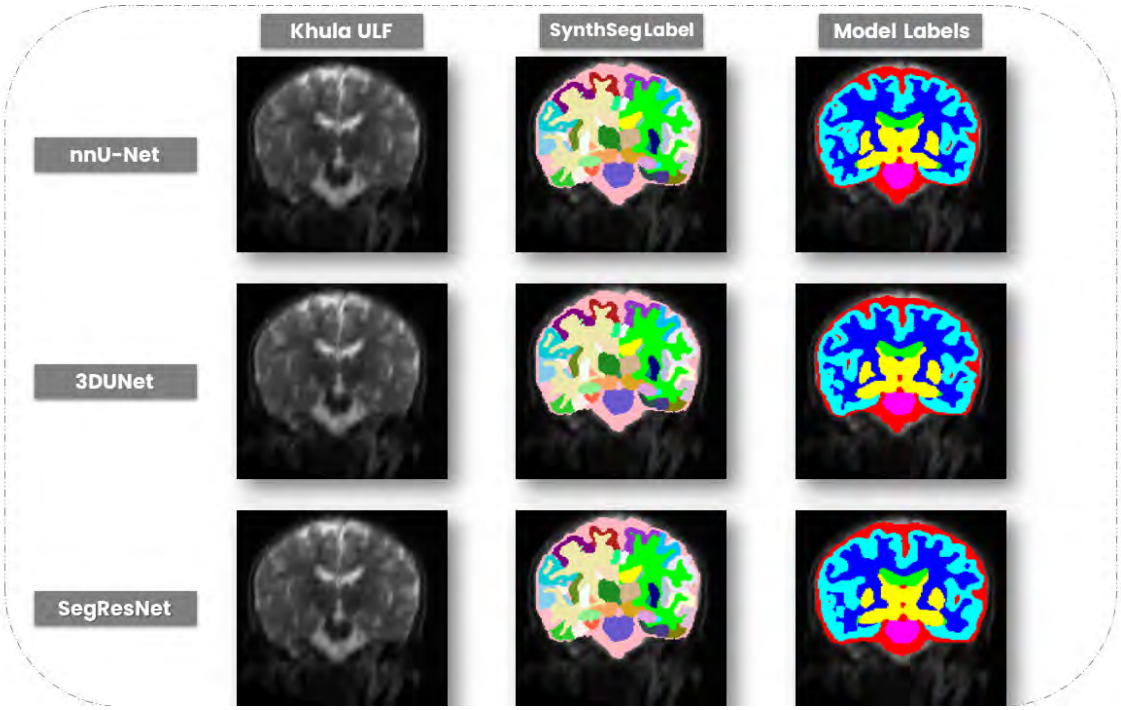



Figure 8: Qualitative Analysis: Comparative Performance of Segmentation Model Predictions on Khula-ULF MRI Samples.

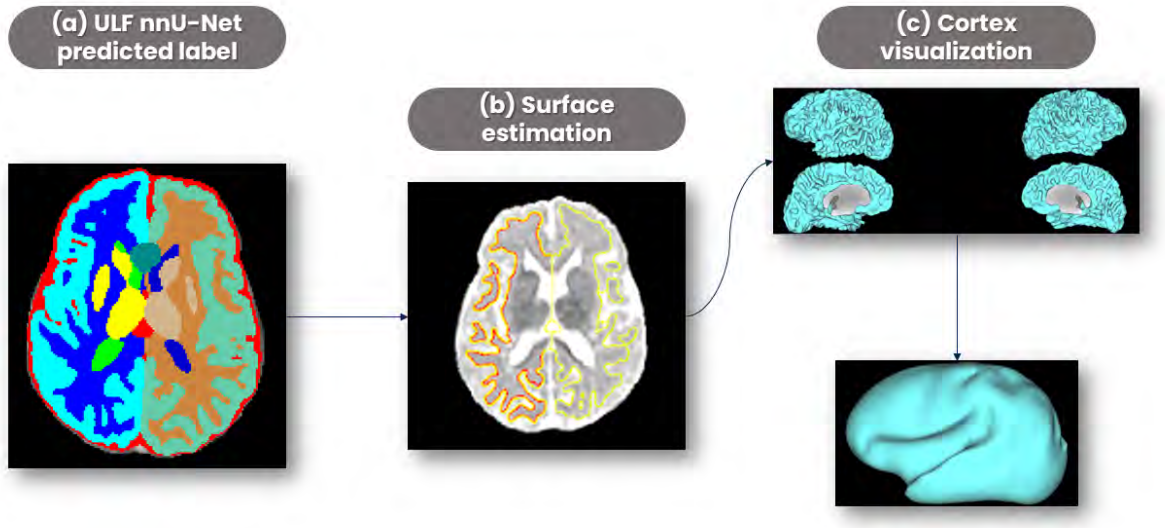



Figure 9: cortical surface extraction workflow. Phase (a): The surface extraction pipeline reconstructs from segmented input image. Phase (b): Surface Estimation: Calculate the boundary between white and grey matter in both hemispheres.; Boundary Refinement: Improve the initial boundary estimate to serve as a base, using feelers to precisely outline the grey matter edge. Phase (c): Surface Visualization: Display edges as surfaces, enabling comprehensive analysis.

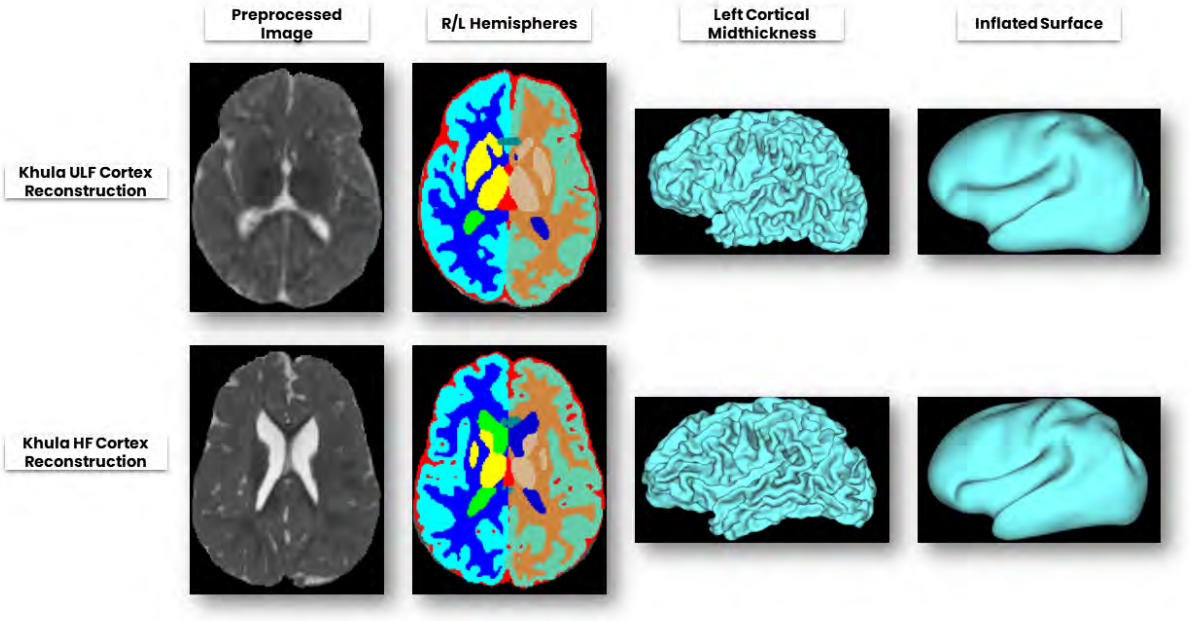



Figure 10: Qualitative results of our extracted surfaces using the dHCP pipeline.

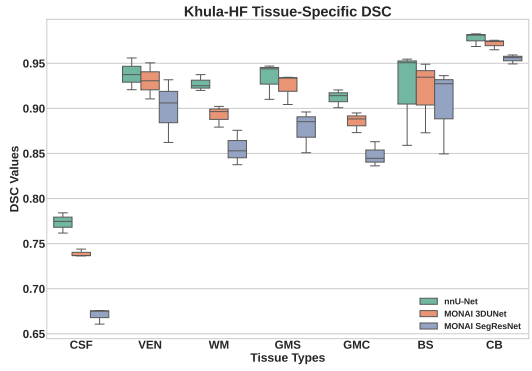



Figure 11: Illustration of DSC values for different brain tissues predicted by various models. Box plots show individual subject variations.

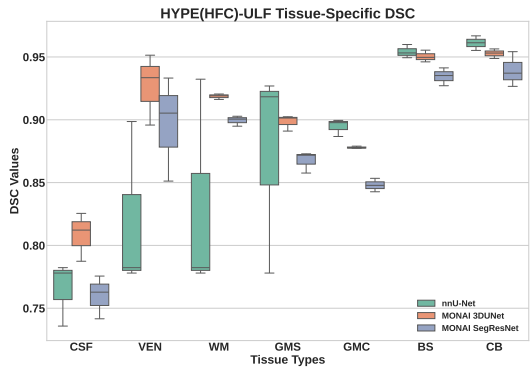



Figure 12: Illustration of DSC values for different brain tissues predicted by various models. Box plots show individual subject variations.

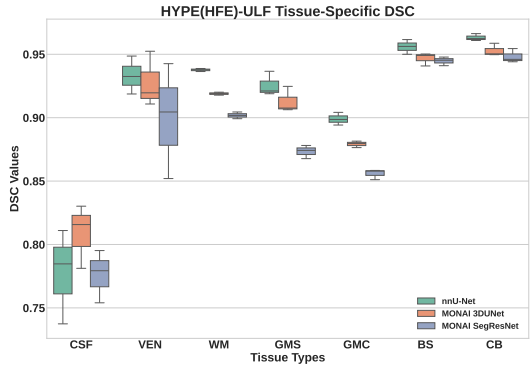



Figure 13: Illustration of DSC values for different brain tissues predicted by various models. Box plots show individual subject variations.

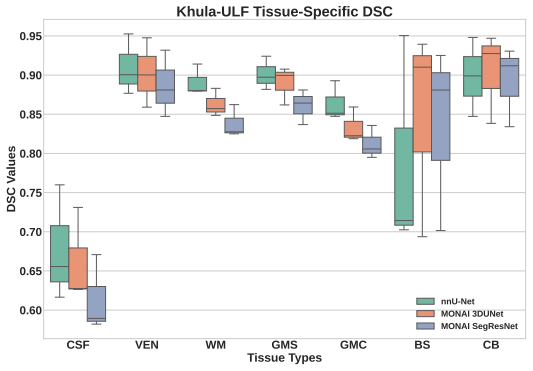



Figure 14: Illustration of DSC values for different brain tissues predicted by various models. Box plots show individual subject variations.

Table 12: Quantitative evaluation of extracted cortical surfaces measured in squared mm.

| DATA TYPE | TOTAL AREA | MEAN AREA |
|---|---|---|
| Khula-HF | **50556.5** | 0.4693979 |
| Khula-ULF | 46526.89 | **0.472892** |

#### *4.1.1. Interpretation of Quantitative Analysis Across Different MRI Data Samples*

The nnU-Net consistently demonstrates superior performance across both HF and ULF MRI data. High Mean DSC values **(Tables 2, 3, 4, 5, 6)** ) indicate exceptional segmentation accuracy, while the lowest Mean HD values confirm excellent boundary detection. Moderate Mean AVD values suggest good volumetric consistency, and high tissue-specific DSC scores **(Tables 7, 8, 9, 10, and 11)**, especially for GMC, validate effective agreement with SynthSeg labels.

The MONAI 3DUNet also performs reliably, with consistently high Mean DSC values across **(Tables 2, 3, 4, 5, and 6)** indicating accurate tissue segmentation. However, its Mean HD varies, showing some sensitivity to MRI field strength. Moderate Mean AVD values reflect decent volumetric accuracy, and consistent GMC-specific DSC values across **(Tables 7, 8, 9, 10, and 11)**, better agreement with SynthSeg label.

Conversely, the MONAI SegResNet shows lower Mean DSC values, indicating less accurate segmentation. High Mean HD values, especially in Khula-HF samples **(Table 3)**, highlight significant edge precision issues. Higher Mean AVD values, particularly in Khula-ULF samples **(Table 6)**, point to poor volumetric accuracy. Lower GMC-specific DSC scores **(Tables 7, 8, 9, 10, and 11)**, reveal challenges in cortex segmentation due to the model's reduced effectiveness in handling noise and low contrast images.

The UNet with ResNet34 backbone, tested on HYPE-HF MRI samples, shows competent Mean DSC **(0.9019, Table 2)** but lags behind other models. High Mean HD **(34.6234, Table 2)**, suggests poor edge detection, while a Mean AVD of 0.0387 **(Table 2)**, indicates reasonable volumetric consistency. Lower GMC-specific DSC scores **(Tables 7, 8, 9, 10, and 11)**, reflect limitations in precise cortex segmentation.

Overall, nnU-Net, leveraging an ensemble approach for inference, emerges as the top performer, offering superior accuracy and consistency across HF and ULF MRI conditions. While MONAI 3DUNet performs well, it shows variability with different field strengths. MONAI SegResNet and UNet with ResNet34 backbone exhibit notable limitations in precision and consistency, particularly under challenging ULF MRI conditions. These findings underscore the importance of selecting models based on specific imaging characteristics and applications.

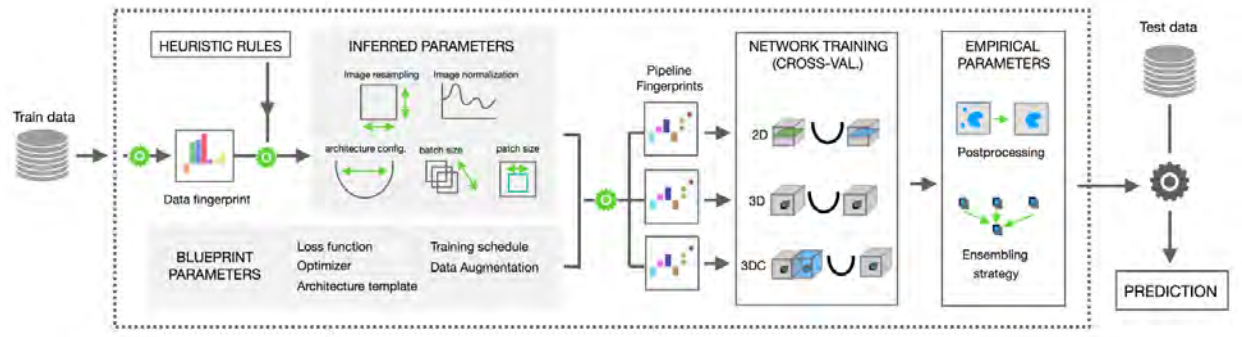



Figure 15: nnU-Net complete workflow and configuration (Isensee et al., 2019).

### *4.1.2. Best Segmentation Model Configuration and Architecture*

The nnU-Net configuration **(shown in Figure 15)** employs a U-Net architecture that handles both 3D and 2D input data **(shown in Figure 16)**. The 2D U-Net starts with two convolutional layers (32 filters, 3x3 size, stride of 1) with InstanceNorm and LeakyReLU activations, followed by max pooling. Encoder blocks double the filters (64, 128, 256, 512), followed by max pooling. The decoder uses up-convolutions and concatenates encoder feature maps, ending with a 1x1 convolutional output layer using softmax activations. The 3D U-Net starts with 32-filter 3D convolutions, increasing through stages (64, 128, 256, 320) with max pooling. Up-convolutions restore dimensions, merging with encoder features via skip connections, concluding with a 1x1x1 convolution for precise volumetric segmentation. Our nnU-Net segmentation framework has been deployed via **MScThesis** with accessibility features for interactive experiment as shown in **Figure 17; appendix**

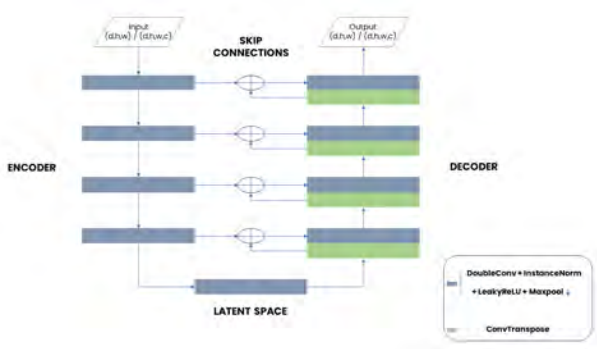



Figure 16: U-Net Architecture for 2D and 3D Segmentation.

## 5. Discussion

Our project demonstrates significant progress in brain cortex segmentation using ULF MRI, particularly beneficial for regions with limited access to HF MRI systems. The integration of AI, especially DL techniques like CNNs and U-Nets, has been transformative in addressing ULF MRI challenges, such as low signal-to-noise ratios and reduced contrast. Experimental results consistently showed nnU-Net outperforming other models across various imaging conditions.

The success of these segmentation models relied heavily on advanced preprocessing techniques. Image resampling, bias field correction (BFC), denoising, and skull stripping were instrumental in enhancing data quality and consistency. These steps ensured optimized input data, despite ULF MRI's low contrast and resolution. Image resampling standardized voxel spacing, BFC corrected uneven intensities, denoising enhanced signal clarity, and skull stripping removed non-brain tissues.

Combining these preprocessing enhancements with robust segmentation algorithms enabled accurate mapping of complex brain structures, particularly the GMC. This groundwork is essential for surface reconstruction, facilitating detailed analyses of cortical thickness, surface area, and gyrification patterns, which are key factors for diagnosing neurodevelopmental conditions.

The integration of the dHCP pipeline further advanced this work by enabling precise cortical surface reconstruction from ULF MRI scans. Utilizing high-quality segmented images from nnU-Net, the dHCP pipeline helped with the extraction of surfaces for neurodevelopmental assessment. Khula-HF had a larger total surface area, indicating a potentially more detailed surface, while Khula-ULF had a slightly higher average surface area per vertex suggesting a more uniform distribution of vertices across the surface **shown in Table 12**.

Our work achieved GMC-specific DSC of (0.8928, 0.8474, and 0.8513 in Khula ULF; 0.8995, 0.8977, and 0.8867 in HYPE HFC; 0.8986, 0.9042, and 0.8942 in HYPE HFE) for the three test samples in ULF MRI segmentation tasks. This demonstrates state-of-the-art results in a field that remains understudied, particularly leveraging only ULF MRI without super-resolution techniques. Our findings build on existing research in ULF MRI studies, such as those by Baljer et al. (2024) and the SynthSR (Iglesias et al., 2022) method used by Cooper et al. (2024), who explored super-resolution techniques for ULF MRI, aligning with our goals of improving image quality and diagnostic accuracy.

### *5.1. Challenges and Future Work*

While our project has made significant strides, several challenges remain, necessitating ongoing refinements to our methodologies. Adapting algorithms to extremely low-resolution images and ensuring consistent performance across different tissue types have been particularly challenging. Additionally, it's crucial to address limitations such as the use of SynthSeg labels as ground truth, which may not fully capture the complexities of the cortex. Despite these challenges, we are actively pursuing avenues for improvement.

Future research efforts will focus on refining segmentation models and preprocessing techniques to enhance performance, especially on both HF and ULF images. We plan to explore contrastive learning techniques, as proposed by (Chen et al., 2020), which leverage sample comparisons to improve feature representation and prediction accuracy. This approach holds promise for enhancing our segmentation results across various image resolutions and tissue types.

Moreover, integrating super-resolution methods, such as those by Baljer et al. (2024) and the SynthSR (Iglesias et al., 2022) method used by Cooper et al. (2024), present an exciting opportunity to enhance the processing of ULF MRI data. By improving image resolution, we aim to overcome some of the challenges associated with low-resolution imaging, potentially enhancing the accuracy of cortical segmentation and furthering our understanding of cortical thickness

and surface area. These advancements not only expand the applications of neuroimaging but also contribute to setting new standards in medical imaging, particularly in regions lacking HF MRI facilities.

## 6. Conclusions

Our proposed method for this project has shown that advanced AI-driven segmentation algorithms can significantly enhance neuroimaging with ULF MRI, particularly in regions without HF MRI systems. By fine-tuning models like nnU-Net and using advanced preprocessing techniques, we overcame the constraints posed by ULF MRI. Our results show accurate segmentation of complex brain structures, with nnU-Net achieving the highest GMC-specific DSC for the 3 test samples. This enables precise surface reconstructions essential for diagnosing neurodevelopmental disorders. Integrating the dHCP pipeline further enhanced our ability to generate detailed cortical surface models from ULF MRI scans. Future work will focus on refining these techniques for greater efficiency and accuracy. This study marks a significant step in merging AI with ULF MRI technology, paving the way for innovations and broader applications in medical imaging.

## Abbreviations

- **ULF** - Ultra-Low-Field
- **MRI** - Magnetic Resonance Imaging
- **AI** - Artificial Intelligence
- **HF** - High-Field
- **DL** - Deep Learning
- **CNNs** - Convolutional Neural Networks
- **CSF** - Cerebrospinal Fluid
- **VEN** - Ventricles
- **WM** - White Matter
- **GMS** - Grey Matter Subcortex
- **GMC** - Grey Matter Cortex
- **BS** - Brainstem
- **CB** - Cerebellum
- **T1w** - T1-weighted
- **T2w** - T2-weighted
- **DSC** - Dice Similarity Coefficient
- **BFC** - Bias-Field Correction
- **HFC** - Centre for Neuroimaging Sciences
- **HFE** - Evelina Newborn Imaging Centre
- **SM3D** - segmentation-models-3D
- **HD** - Hausdorff Distance
- **AVD** - Average Volumetric Difference
- **QC** - quality control

## Acknowledgments

I would like to express thanks to Dr. Emma C. Robinson, Dr. František Váša, Levente Baljer and Alena Uus for their exceptional supervision and support throughout this research project. I'm also grateful for the skills and knowledge gained from the MAIA MSc. program that enabled me to contribute to this pioneering work.

## References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al., 2016. {TensorFlow}: a system for {Large-Scale} machine learning, in: 12th USENIX symposium on operating systems design and implementation (OSDI 16), pp. 265–283.

Abate, F., Adu-Amankwah, A., Ae-Ngibise, K., Agbokey, F., Agyemang, V., Agyemang, C., Akgun, C., Ametepe, J., Arichi, T., Asante, K., et al., 2024. Unity: A low-field magnetic resonance neuroimaging initiative to characterize neurodevelopment in low and middle-income settings. Developmental Cognitive Neuroscience , 101397.

Al-Amri, S.S., Kalyankar, N.V., et al., 2010. Image segmentation by using threshold techniques. arXiv preprint arXiv:1005.4020 .

Arnold, T.C., Freeman, C.W., Litt, B., Stein, J.M., 2023. Low-field mri: clinical promise and challenges. Journal of Magnetic Resonance Imaging 57, 25–44.

Baljer, L., Zhang, Y., Bourke, N.J., Donald, K.A., Bradford, L.E., Ringshaw, J.E., Williams, S.R., Deoni, S.C., Williams, S.C., Team, K.S.S., et al., 2024. Multi-orientation u-net for super-resolution of ultra-low-field paediatric mri. bioRxiv , 2024–02.

Baniasadi, M., Petersen, M.V., Gonçalves, J., Horn, A., Vlasov, V., Hertel, F., Husch, A., 2023. Dbsegment: Fast and robust segmentation of deep brain structures considering domain generalization. Human Brain Mapping 44, 762–778.

Bertels, J., Eelbode, T., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R., Blaschko, M.B., 2019. Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice, 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. 92–100.

Billot, B., Magdamo, C., Cheng, Y., Arnold, S.E., Das, S., Iglesias, J.E., 2023. Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain mri datasets. Proceedings of the National Academy of Sciences 120, e2216399120.

Buxton, R.B., Edelman, R.R., Rosen, B.R., Wismer, G.L., Brady, T.J., 1987. Contrast in rapid mr imaging: T1-and t2-weighted imaging. J Comput Assist Tomogr 11, 7–16.

Cabezas, M., Oliver, A., Lladó, X., Freixenet, J., Cuadra, M.B., 2011. A review of atlas-based segmentation for magnetic resonance brain images. Computer methods and programs in biomedicine 104, e158–e177.

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 .

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.

Choi, R.Y., Coyner, A.S., Kalpathy-Cramer, J., Chiang, M.F., Campbell, J.P., 2020. Introduction to machine learning, neural networks, and deep learning. Translational vision science & technology 9, 14–14.

Cooper, R.E., Hayes, R., Arnold, T.C., Stein, J., Jalbrzikowski, M., 2024. Bridging the gap: improving correspondence between low-field and high-field magnetic resonance images in young people. original research article. Frontiers in Neurology 15, 1339223.

Dehdasht-Heydari, R., Gholami, S., 2019. Automatic seeded region growing (asrg) using genetic algorithm for brain mri segmentation. Wireless Personal Communications 109, 897–908.

Ertl-Wagner, B., Wagner, M., 2023. Ultralow-field-strength mri and artificial intelligence: How low can we go and how high can we reach?

Fernández-Pena, A., 2023. Surface-Based Tools for Characterizing the Human Brain Cortical Morphology. Ph.D. thesis.

Fischl, B., 2012. Freesurfer. Neuroimage 62, 774–781.

Hamghalam, M., Wang, T., Lei, B., 2020. High tissue contrast image synthesis via multistage attention-gan: application to segmenting brain mr scans. Neural Networks 132, 43–52.

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.

Holmes, J., Sacchi, L., Bellazzi, R., et al., 2004. Artificial intelligence in medicine. Ann R Coll Surg Engl 86, 334–8.

Hoopes, A., Mora, J.S., Dalca, A.V., Fischl, B., Hoffmann, M., 2022. Synthstrip: skull-stripping for any brain image. NeuroImage 260, 119474.

Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J., 1993. Comparing images using the hausdorff distance. IEEE Transactions on pattern analysis and machine intelligence 15, 850–863.

Iglesias, J.E., Schleicher, R., Laguna, S., Billot, B., Schaefer, P., McKaig, B., Goldstein, J.N., Sheth, K.N., Rosen, M.S., Kimberly, W.T., 2022. Accurate super-resolution low-field brain mri. arXiv preprint arXiv:2202.03564 .

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.

Isensee, F., Jäger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2019. Automated design of deep learning methods for biomedical image segmentation. arXiv preprint arXiv:1904.08128 .

Islam, K.T., Zhong, S., Zakavi, P., Chen, Z., Kavnoudias, H., Farquharson, S., Durbridge, G., Barth, M., McMahon, K.L., Parizel, P.M., et al., 2023. Improving portable low-field mri image quality through image-to-image translation using paired low-and high-field images. Scientific Reports 13, 21183.

Jacob, J., Ciccarelli, O., Barkhof, F., Alexander, D.C., 2021. Disentangling human error from the ground truth in segmentation of medical images. Advances in Neural Information Processing Systems 33, 15750–15762.

Jalab, H.A., Hasan, A.M., 2019. Magnetic resonance imaging segmentation techniques of brain tumors: A review. Archives of Neuroscience 6.

Javed, K., Reddy, V., Lui, F., 2023. Neuroanatomy, cerebral cortex, in: StatPearls [Internet]. StatPearls Publishing.

Jin, H., Song, Q., Hu, X., 2019. Auto-keras: An efficient neural architecture search system, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 1946–1956.

Keith, R., 2023. Radiology's role in neurological disorder management: Imaging techniques and diagnostic advancements. Journal Environmental Sciences And Technology 2, 198–207.

Kwon, G.R., Basukala, D., Lee, S.W., Lee, K.H., Kang, M., 2016. Brain image segmentation using a combination of expectation-maximization algorithm and watershed transform. International Journal of Imaging Systems and Technology 26, 225–232.

Langen, K.J., Galldiks, N., Hattingen, E., Shah, N.J., 2017. Advances in neuro-oncology imaging. Nature Reviews Neurology 13, 279–289.

Lee, B., Yamanakkanavar, N., Choi, J.Y., 2020. Automatic segmentation of brain mri using a novel patch-wise u-net deep architecture. Plos one 15, e0236493.

Liu, Y., Leong, A.T., Zhao, Y., Xiao, L., Mak, H.K., Tsang, A.C.O., Lau, G.K., Leung, G.K., Wu, E.X., 2021. A low-cost and shielding-free ultra-low-field brain mri scanner. Nature communications 12, 7238.

Makropoulos, A., Robinson, E.C., Schuh, A., Wright, R., Fitzgibbon, S., Bozek, J., Counsell, S.J., Steinweg, J., Vecchiato, K., Passerat-Palmbach, J., et al., 2018. The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction. Neuroimage 173, 88–112.

McKinley, R., Wepfer, R., Aschwanden, F., Grunder, L., Muri, R., Rummel, C., Verma, R., Weisstanner, C., Reyes, M., Salmen, A., et al., 2021. Simultaneous lesion and brain segmentation in multiple sclerosis using deep neural networks. Scientific reports 11, 1087.

Molnár, Z., Clowry, G.J., Šestan, N., Alzu'bi, A., Bakken, T., Hevner, R.F., Hüppi, P.S., Kostović, I., Rakic, P., Anton, E., et al., 2019. New insights into the development of the human cerebral cortex. Journal of anatomy 235, 432–451.

Ogbole, G.I., Adeyomoye, A.O., Badu-Peprah, A., Mensah, Y., Nzeh, D.A., 2018. Survey of magnetic resonance imaging availability in west africa. Pan African Medical Journal 30.

O'shea, K., Nash, R., 2015. An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458 .

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.

Ramzan, F., Khan, M.U.G., Iqbal, S., Saba, T., Rehman, A., 2020. Volumetric segmentation of brain regions from mri scans using 3d convolutional neural networks. IEEE Access 8, 103697–103709.

Rayed, M.E., Islam, S.S., Niha, S.I., Jim, J.R., Kabir, M.M., Mridha, M., 2024. Deep learning for medical image segmentation: State-of-the-art advancements and challenges. Informatics in Medicine Unlocked , 101504.

Ren, J., Hu, Q., Wang, W., Zhang, W., Hubbard, C.S., Zhang, P., An, N., Zhou, Y., Dahmani, L., Wang, D., et al., 2022. Fast cortical surface reconstruction from mri using deep learning. Brain informatics 9, 6.

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.

Schuh, A., Makropoulos, A., Wright, R., Robinson, E.C., Tusor, N., Steinweg, J., Hughes, E., Grande, L.C., Price, A., Hutter, J., et al., 2017. A deformable model for the reconstruction of the neonatal cortex, in: 2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017), IEEE. pp. 800–803.

Shen, D., Wu, G., Suk, H.I., 2017. Deep learning in medical image analysis. Annual review of biomedical engineering 19, 221–248.

Shetewi, S.G., Al Mutairi, B.S., Bafaraj, S.M., 2020. The role of imaging in examining neurological disorders; assessing brain, stroke, and neurological disorders using ct and mri imaging. Advances in Computed Tomography 9, 1–11.

Shoghli, A., Chow, D., Kuoy, E., Yaghmai, V., 2023. Current role of portable mri in diagnosis of acute neurological conditions. Frontiers in Neurology 14, 1255858.

Singh, N.M., Harrod, J.B., Subramanian, S., Robinson, M., Chang, K., Cetin-Karayumak, S., Dalca, A.V., Eickhoff, S., Fox, M., Franke, L., et al., 2022. How machine learning is powering neuroimaging to improve brain health. Neuroinformatics 20, 943–964.

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.

Solovyev, R., Kalinin, A.A., Gabruseva, T., 2022. 3d convolutional neural networks for stalled brain capillary detection. Computers in biology and medicine 141, 105089.

Stiles, J., Jernigan, T.L., 2010. The basics of brain development. Neuropsychology review 20, 327–348.

Tocchio, S., Kline-Fath, B., Kanal, E., Schmithorst, V.J., Panigrahy, A., 2015. Mri evaluation and safety in the developing brain, in: Seminars in perinatology, Elsevier. pp. 73–104.

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.

Tustison, N.J., Cook, P.A., Holbrook, A.J., Johnson, H.J., Muschelli, J., Devenyi, G.A., Duda, J.T., Das, S.R., Cullen, N.C., Gillen, D.L., et al., 2021. The antsx ecosystem for quantitative biological and medical imaging. Scientific reports 11, 9068.

Uus, A.U., Kyriakopoulou, V., Makropoulos, A., Fukami-Gartner, A., Cromb, D., Davidson, A., Cordero-Grande, L., Price, A.N., Grigorescu, I., Williams, L.Z., et al., 2023. Bounti: Brain volumetry and automated parcellation for 3d fetal mri. bioRxiv .

Vasung, L., Turk, E.A., Ferradal, S.L., Sutin, J., Stout, J.N., Ahtam, B., Lin, P.Y., Grant, P.E., 2019. Exploring early human brain development with structural and physiological neuroimaging. Neuroimage 187, 226–254.

Weese, J., Lorenz, C., 2016. Four challenges in medical image analy-

sis from an industrial perspective.

Yaniv, Z., Lowekamp, B.C., Johnson, H.J., Beare, R., 2018. Simpleitk image-analysis notebooks: a collaborative environment for education and reproducible research. Journal of digital imaging 31, 290–303.

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, Z., 2018. Improved adam optimizer for deep neural networks, in: 2018 IEEE/ACM 26th international symposium on quality of service (IWQoS), Ieee. pp. 1–2.

Zieff, M.R., Miles, M., Mbale, E., Eastman, E., Ginnell, L., Williams, S.C., Jones, D.K., Alexander, D.C., Wijeratne, P.A., Gabard-Durnam, L.J., et al., 2024. Characterizing developing executive functions in the first 1000 days in south africa and malawi: The khula study. Wellcome Open Research 9, 157.

## 7. Appendix

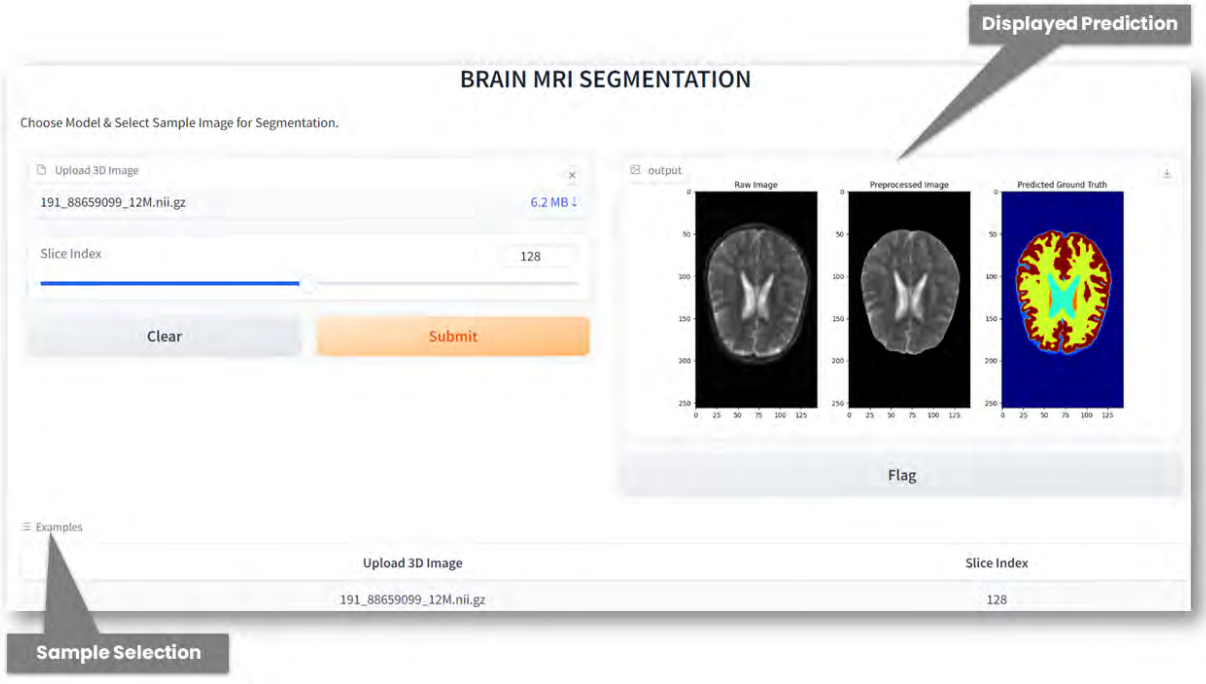
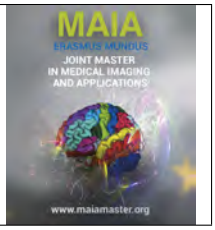



Figure 17: A user interactive surface of our deployed nnU-Net segmentation framework. You can access the complete project code *here*.

Medical Imaging and Applications

Master Thesis, July 2024




# Comparison and evaluation of finite element analysis and deep learning methods for breast biomechanical models

Hadeel Awwad, Eloy García Marcos, Robert Martí Marly

*Research institute of Computer Vision and Robotics (ViCOROB), Universitat de Girona, Catalonia, Spain*

**Abstract**

Accurate breast compression simulation is essential in medical imaging, mainly in mammography. We compare and evaluate two established Finite Element Methods (FEM) for this purpose: NiftySim, a previously developed framework for generating compressed breast phantoms using a biomechanical finite element (FE) model from Breast CT (BCT) volumes, and FEBio, a specialized non-linear solver for biomechanics applications. The rising computational cost of traditional FEM limits their clinical utility. This work also explores the use of a deep learning framework, Physics-based Graph Neural Networks (PhysGNN), as a data-driven alternative for breast compression simulation. While PhysGNN has been used for data-driven modeling in other domains, this thesis presents the first investigation of their potential in this specific context. Unlike conventional data-driven models, PhysGNN incorporates valuable mesh structural information and facilitates inductive learning on unstructured grids, making it well-suited for capturing the complex geometries of breast tissue. The model is trained on deformations obtained from incremental FEM simulations, and its performance is assessed by comparing the predicted nodal positions (using 3D Euclidean distance) with those extracted from the FE simulations. Our empirical evaluation demonstrates that NiftySim and FEBio are capable of achieving a more realistic compression of breast phantoms compared to previously published BCT simulations. FEBio simulations yielded approximations of mammographic deformation closer to those obtained with NiftySim. However, NiftySim offers superior accuracy at the expense of higher computational cost. The investigated deep learning architecture, PhysGNN, shows promise for achieving accurate and rapid approximations of breast deformation during compression. Its potential for enhanced computational efficiency makes it a suitable candidate for real-world breast compression scenarios.

*Keywords:* Finite Element Analysis, CT, Data-Driven model, Graph Neural Netowrk, Meshing

## 1. Introduction

Breast biomechanical models are crucial tools for simulating tissue deformations using Finite Element Analysis (FEA). This simulation capability aids clinicians and medical device manufacturers in addressing challenges encountered during various breast procedures. Personalized models, incorporating individual breast geometry, mechanical properties, and boundary conditions, are essential for reliable predictions. These models have found valuable applications in breast augmentation, image-guided surgery, and tumor tracking.

In this thesis, we focus on simulating breast compression using FEM. We employ two established FEM: NiftySim (Johnsen et al., 2015) and FEBio (Maas et al., 2012).

While the FEM has established itself as a reliable technique for approximating soft tissue deformation, it inherently faces a trade-off between achieving high accuracy and maintaining efficient computational speed.

Therefore, in this thesis, we also propose to explore a novel data-driven model named PhysGNN, originally designed for capturing intraoperative brain shift in image-guided neurosurgery (Salehi and Giannacopoulos, 2022). PhysGNN utilizes Graph Neural Networks (GNNs) to exploit the structural information of the FE mesh, including connectivity and distances between nodes. This structural awareness allows PhysGNN to accurately approximate tissue deformation, traditionally

handled by FEM but often facing limitations in computational speed. Notably, PhysGNN offers faster computation compared to FEM. To our knowledge, this is the first application of GNNs for approximating breast tissue behavior under compression. PhysGNN's framework integrates seamlessly with traditional FEM workflows, potentially enhancing simulation speed without compromising accuracy. This paves the way for applications demanding both precision and efficiency. To assess the performance of the PhysGNN model, we employed two strategies: hold-out and leave-one-deformation-out. These strategies differ in how the data is split for training, validation, and testing. We compared the results obtained with PhysGNN to those from FEA, which served as our baseline. In the hold-out experiment, PhysGNN demonstrated the best performance, achieving a high degree of agreement with the ground truth data. It achieved an absolute displacement error of less than 1 mm for 99.96% of the nodes. Additionally, PhysGNN successfully predicted breast deformation under prescribed forces in 0.01 ± 0.06 seconds on a GPU and 0.42 ± 0.04 seconds on a CPU, compared to 300 and 6634 seconds required by traditional FE methods.

An additional aim was also to develop a more simplified and homogenous framework in terms of programming language and development platform compared to existing approaches. Hence we implement the biomechanical modeling using Python, unlike other works that employ different programming languages for various stages. This simplifies the workflow and streamlines the process.

## 2. State of the art

### *2.1. Finite element methods and application to biomechanical breast models*

FEM discretizes an object to approximately solve the differential equations describing the physical conditions (Pasciak, 1995). The boundary conditions are considered as input from which the algorithm approximates to the corresponding solution (Pasciak, 1995). The continuum problem is approximated by a method where the continuum is partitioned into a finite number of elements and a finite set of parameters determines the performance of these elements. The solution of the whole system as a set of its elements pursues exactly the same principles that apply to standard discrete problems (Zienkiewicz et al., 2005). Biomechanical modeling (BM) of the behavior of anatomical structures under different loads is a necessary step for numerous academic and clinical applications. Corresponding partial differential equations (PDEs) control the physical phenomenon being modeled such as the deformation of organs like the liver, prostate, stomach, breast, and other virtual organs in augmented reality applications (Phellan et al., 2021a). FEM specifically for modeling the mechanical response of breast tissue has been used for several applications with different complexities of biomechanical models and materials (García et al., 2018; Hipwell et al., 2016). More specifically, the mechanical response of breast tissue under mammographic compression has been modeled using FEM. An overview of the tissues, software packages, computational time, and type of studies used is presented in Table 1).

Although FEM have demonstrated success in registering multimodal breast images, their computational demands can hinder their practical use in clinical settings and integration into routine workflows. Additionally, some approaches necessitate iterative adjustments to various parameters, like compression thickness, material properties, and breast rotation, to account for inherent uncertainties present in clinical data (referred to as "Optim." for optimization in Table 1). The repetitive nature of FEM simulations, often requiring numerous computations, significantly increases the overall processing time. Consequently, achieving accurate and real-time modeling of soft tissue deformation remains a significant challenge. To achieve real-time compatibility, various strategies have been explored to lessen the computational burden associated with FEM. Notably, some approaches have targeted improvements to linear solvers, a known bottleneck within the FEM workflow (Mendizabal et al., 2020). Notably, (Han et al., 2013) explored a GPU implementation that relies on the Total Lagrangian Explicit Dynamics (TLED) formulation proposed by (Miller et al., 2007). This TLED formulation is considered highly suitable for modeling breast biomechanics (Mendizabal et al., 2020). Moreover, NiftySim is a GPU-based solver, used for simulating soft tissue deformations. While its applications extend beyond breast modeling, it proved to be valuable in our study. Simulation Open Framework Architecture (SOFA) represents another approach to accelerating computations for soft tissue simulations. SOFA, known for its effectiveness with GPU-based solvers, has been successfully applied to studies involving prostate deformation, as referenced in (Moreira et al., 2013).

### *2.2. Machine and deep learning algorithms*

Machine learning algorithms have emerged as a powerful tool for predicting the mechanical behavior of various anatomical structures (Phellan et al., 2021b). Machine learning models excel in their ability to make real-time predictions after a preliminary training phase. These models operate by learning complex relationships from data, allowing them to forecast outcomes efficiently once trained.

Mendizabal et al. (2020) explored simulating ultrasound image deformations during an ultrasound-guided breast biopsy. The researchers employed a U-net architecture trained on a limited synthetic dataset. Their primary focus was establishing a correlation between the

Table 1: Literature review of finite element methods for breast deformation

| Reference | Application | Tissues | Computation Time | Software | Studies |
|---|---|---|---|---|---|
| Azar et al. (2000) | MR image-guided biopsy | Adipose, glandular and lesion | < 30 mins | ABAQUS | Clinical |
| Samani et al. (2001) | Breast compression | Adipose, glandular | - | ABAQUS | Phantom |
| Ruiter et al. (2002) | Cancer diagnosis | - | - | ANSYS | Clinical |
| Ruiter et al. (2006) | Image registration | Adipose, glandular | - | ANSYS | Clinical |
| Tanner et al. (2006) | Breast compression | Adipose, glandular | - | ANSYS | Clinical |
| Chung et al. (2008) | Image registration | Breast volume | - | CMISS | Phantom |
| Hsu et al. (2011) | Breast compression (phantom generation) | Adipose, glandular and skin | 3 - 4 h | LS-DYNA | Phantom |
| Han et al. (2011) | Breast compression | Adipose, glandular and tumor | 312 mins (Explicit) | ABAQUS | Clinical |
| Hopp et al. (2013) | Image registration | Breast volume | 20 mins (Optim.:120 mins) | ABAQUS | Clinical |
| Lee et al. (2013) | Image registration | Breast volume | - | CMISS | Clinical |
| Mertzanidou et al. (2014) | Image registration | Breast volume | 2 h | ITK | Clinical |
| Sturgeon et al. (2016) | Image registration | Adipose, glandular and skin | 2 h 13 mins | FEBio | Phantom |
| Liu et al. (2017) | Simulation compression | Adipose, glandular | - | ABAQUS | Clinical |
| Martínez-Martínez et al. (2017) | Simulation compression | Adipose, glandular and skin | - | ANSYS | Clinical |
| García et al. (2019) | Image registration | Adipose, glandular | 61 min (Optim.) | NiftySim | Clinical |

partial surface deformation observed under the ultrasound probe and the internal breast deformation. This approach achieved real-time prediction of lesion displacement with good accuracy. However, limitations existed. The model did not account for the natural variations (heterogeneity) within real breast tissue nor the complex boundary conditions encountered in clinical settings. Additionally, its reliance solely on surface displacement data limited its sensitivity to patient-specific elasticity. The researchers evaluated their model's performance by comparing its predicted lesion displacements with those generated by a high-fidelity FEM sim-

ulation (considered ground truth). They used the mean error metric to quantify this comparison. Additionally, since accurate lesion prediction was crucial, they employed the Target Registration Error (TRE) to assess the difference between their predicted lesion location and the actual position within the phantom breast. The model was tested with varying probe displacements ranging from less than 12.5 mm to over 27.5 mm, achieving a mean TRE between 2.7 mm and 5.8 mm, respectively. However, a significant limitation of their approach was the need to retrain the model for each new breast geometry. This restricted its applicability to specific types and numbers of compression tools, such as various probe shapes (Mendizabal et al., 2020).

Building on MRI-derived non-linear finite element models, previous research by (Martínez-Martínez et al., 2017) and (Rupérez et al., 2018) explored modeling the mechanical response of breast tissue under mammographic compression. Their primary objective was to achieve faster multimodal registration and simulate breast tissue behavior during image-guided procedures like biopsies. They investigated three machine learning models for this purpose: decision trees (DT), extremely randomized trees (ERT), and random forests (RF). Initial experiments were conducted using phantoms, followed by extension to clinical datasets. For the evaluation, they calculated the mean 3D Euclidean distance between the nodes predicted by the models and the nodes extracted from the FE simulation, which served as the ground truth. Their experiments revealed that ERT outperformed the other models, achieving an average error of only 0.62 mm. However, the study was limited by the relatively small dataset used for evaluation (10 phantoms and 10 clinical cases), potentially hindering the model's generalizability. Additionally, the model only considered a 20% compression ratio, which is significantly lower than compression levels typically used in mammography. Finally, the model simplified the breast tissue composition by representing it with just three types: fatty, glandular, and skin.

Building upon a previously developed biomechanical model (Hopp et al., 2012, 2013) that incorporated four tissue types (fatty, glandular, muscular, and skin) and unloaded breast state estimation, (Said et al., 2023) compared different machine learning models for predicting breast deformation under realistic mammographic compression. They investigated three models: ERT, Extreme Gradient Boosting (XGBoost), and a deep learning model called Attention-Based Bidirectional Long Short-Term Memory (Att-BLSTM). They evaluated their machine learning (ML) and deep learning (DL) models using data from 516 breasts. Biomechanical models were automatically generated from T2-weighted MR images for each breast. The models were trained and tested using a FEM simulation that mimicked mammographic compression at varying ratios (up to 76% based on patient metadata). Additionally, the researchers analyzed the models' performance to factors like compression ratio, tissue type accuracy, and overall breast volume. The authors evaluated their models by comparing their predicted node positions with the deformed nodes from their FEM simulation as the ground truth. They calculated the Root Mean Square Error (RMSE) for each node and then averaged it across all nodes in a breast (one dataset). Finally, they reported the mean and median RMSE across all datasets in the validation set. Said et al. (2023) reported that their models achieved an average prediction error of 4.7 millimeters (mean RMSE) and a median error of 3.4 millimeters across 516 breasts. Notably, these models offered a significant speedup of roughly 240 times compared to the original FEM model simulation.

Current machine and deep learning approaches often overlook valuable information embedded within the finite element (FE) mesh structure. This information includes details about how nodes are connected and the distances between them. To address this limitation, we propose adapting a well-established approach called PhysGNN to the breast compression problem. PhysGNN leverages Graph Neural Networks, which are specifically designed to exploit the inherent structural information present in graphs. In the context of our work, this translates to effectively utilizing the connectivity and distances between nodes within the FE mesh for improved breast tissue deformation prediction. This capability not only allows PhysGNN to incorporate crucial information about the mesh structure but also enhances computational efficiency when learning from high-quality, high-node-count meshes. This efficiency stems from the message-passing framework employed by GNNs.

## 3. Material and methods

This thesis work can be summarized in the following workflow (refer to Figure 1 for a visual representation):

1. **Comparison of Traditional FEM:** We compared the performance of two established FEM approaches, NiftySim and FEBio, in simulating breast compression during mammography. While traditionally FEM simulations have served as the gold standard for this purpose, we explored the potential of a data-driven alternative.
2. **Ground Truth and Mesh Features Extraction:** Specifically, we utilized the nodal displacements obtained from NiftySim incremental simulations of a single FE mesh model as the ground truth data for training a deep learning model called PhysGNN. Having ground truth data from just one reliable FEM method was sufficient to train the model effectively. The geometric information from the input breast FE mesh consists of an adjacency matrix and a set of edge weights:

- The adjacency matrix is a binary matrix describing the connections between nodes.
- The edge weights are calculated as the inverse of the Euclidean distance between nodes, providing the network with spatial information.

3. **PhysGNN for Structural Information:** To leverage the structural information within the FE mesh, we employed a Graph Neural Network. GNNs are adept at handling data structured as graphs, where nodes represent points of interest and connections between them encode relationships. In our case, the FE mesh itself forms the graph, with vertices corresponding to nodes and the edges corresponding to the connections between nodes within an element (Pfaff et al., 2020). Through a process called message passing, GNNs enable nodes to exchange information with their neighbors, considering both their features and the connections. This allows the adapted PhysGNN model to capture the inherent structural information of the FE mesh, which is critical for the accurate prediction of breast tissue deformation under compression. Notably, GNNs share parameters across the entire mesh, that ensures learning a constant number of parameters independent of the mesh size and promoting efficient learning, and enabling generalization to unseen breast geometries.

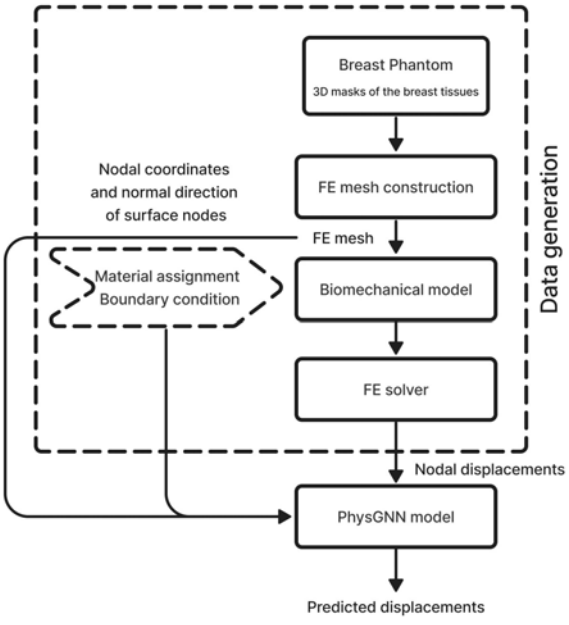



Figure 1: Flow chart of the proposed method to simulate the breast compression in mammography.

### *3.1. Breast phantoms*

Computational breast models are increasingly utilized in breast imaging research to assess and enhance new imaging systems and methods (Bakic et al., 2011; Hsu et al., 2013; Li et al., 2009; Segars et al., 2014). They offer a practical solution for conducting studies that would be challenging or unfeasible with human participants due to high costs or safety risks. To optimize and compare different imaging modalities effectively, these phantoms must accurately simulate the breast in various positions and compression states required by the modalities. For instance, dedicated Breast CT (BCT) captures images of the breast in the prone position without compression, whereas mammography and tomosynthesis involve the patient standing with the breast compressed.

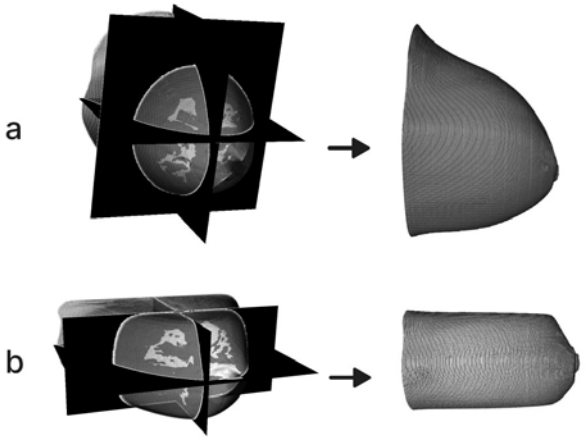



Figure 2: Phantom 3 in the dataset (Sarno et al., 2021a,b). Uncompressed geometry (top) and its corresponding compressed geometry (bottom) after using compression software (Zyganitidis et al., 2007).

The dataset of breast phantoms was accessed from Zenodo public storage, available as an Open Access database (Sarno et al., 2021c). This dataset includes two separate collections: one with 150 uncompressed phantoms (for breast computed tomography studies (Sarno et al., 2021a)) and another with 60 compressed phantoms (for digital breast tomosynthesis DBT and digital mammography DM studies (Sarno et al., 2021b)). The files are stored as DICOM files, where voxel values represent different materials: 0 for air, 1 for adipose tissue, 2 for glandular tissue, and 3 for skin tissue. Additionally, a datasheet accompanies the folders, listing the pixel pitch size and slice thickness for each phantom. The phantoms are named sequentially, and both compressed and uncompressed versions corresponding to the same clinical acquisition share the same name.

The computational breast phantoms were created from clinical breast images previously acquired at UC Davis (California, USA) using an in-house developed BCT scanner. This project used 150 breast volume datasets from 150 different patients. The images were captured with a first-generation BCT scanner (Boone et al., 2001) operating at 80 kV with Cu filtration. The

flat panel detector used 2x2 binning mode with a pixel pitch of 0.388 mm (native pixel pitch of 0.194 mm). The scan protocols involved 500 projections with a complete 360° gantry rotation. The voxel size in the reconstructed coronal slices varied from 0.1938 mm to 0.4274 mm in-plane and from 0.1907 mm to 0.2375 mm in the axial direction (slice thickness). The raw CT data, reconstructed using the FDK algorithm, were corrected for cone beam artifacts, geometric distortion, and cupping artifacts. All BCT examinations were performed with a mean glandular dose of 5 mGy. The uncompressed computational breast phantoms were created from these clinical images using a semi-automatic tissue classification algorithm (Mettivier et al., 2020). This algorithm, developed in Matlab R2019a utilizing routines from the Segmentation toolkit, categorizes each image voxel into one of four material types: adipose tissue, fibroglandular tissue, skin, and air.

The compressed versions of these phantoms, intended for in silico investigations in DBT and DM, were generated from the uncompressed versions using specific software described in (Zyganitidis et al., 2007) as shown in Figure 2. This software computes the compressed version of the pendant breast model based on Young's modulus of elasticity of the materials and the final compressed thickness, which is extracted from the DICOM header information of the original DM examinations. This compression software has been extensively described and used in previous dosimetry and imaging studies in 2D digital mammography. The average compressed thickness for the cohort of compressed computational phantoms was 61 mm. This compression process allows the same "digital patient" to be used in both uncompressed geometry for BCT simulated exams and compressed geometry for DM/DBT exams.

### *3.2. Biomechanical breast model*

The biomechanical breast model is based on the method of generating compressed breast phantoms using a biomechanical finite element model from BCT volumes developed in the work of (García et al., 2020), by simulating physically realistic tissue deformation. It estimates a configuration of the breast comparable to its shape in mammography or breast tomosynthesis based on the breast geometry observed with BCT in 3D. Corresponding mammograms or compressed image data were not available for the BCT dataset used in our study. So, we used the corresponding compressed phantoms as a reference for the compression thickness of each phantom and ground truth. An overview of generating a biomechanical breast model is shown in Figure 3

The biomechanical breast model used for FEBio aims to mimic the (García et al., 2020) model, but with slight adjustments that will be explained in the following sections. These adjustments are necessary to ensure the model runs to completion.

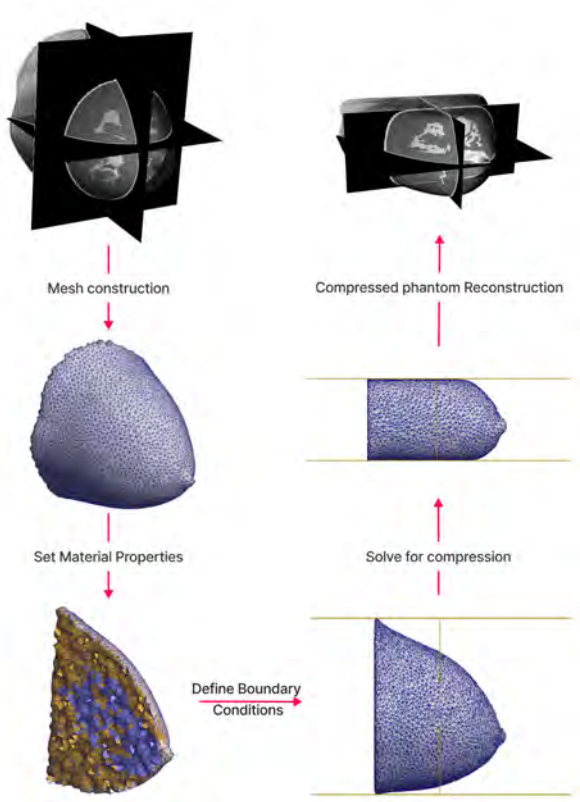



Figure 3: Overview of the steps to set up and solve for the compressed breast geometry using finite-element analysis. Once the FE analysis is complete, the deformation is applied to the phantom to reposition and compress the breast.

#### *3.2.1. Mesh Construction*

To acquire patient-specific breast geometry, the DICOM series for each phantom were converted into NRRD images. These preclassified voxelized images underwent resampling using Nearest Neighbor Interpolater to achieve voxel spacing of 0.273 $mm^3$, ensuring isotropic voxels, crucial for generating a high-quality mesh for FEA. The choice of 0.273 $mm^3$ cubic resolution was motivated by its negligible tissue loss in the compressed phantom reconstruction compared to the original uncompressed phantom, as referenced in (García et al., 2020).

To generate a digital representation of the breast anatomy, a meshing technique was employed. This technique involves discretizing the breast volume into a collection of small, interconnected elements. These elements, often referred to as mesh elements or simply elements, collectively form a net-like structure that approximates the complex geometry of the breast. Unlike a traditional net with squares, this mesh utilizes tetrahedrons, which are three-dimensional shapes with four triangular faces and four nodes (points). This process of creating the mesh from a 3D object is called meshing. The construction of the model geometry from

resampled and preclassified breast images employed a tetrahedral meshing approach. We utilized Pygalmesh (Schlömer), a Python interface to the CGAL library (The CGAL Project, 2024) to generate high-quality 3D volume meshes for our breast models. Pygalmesh offers various parameters to control the mesh properties. We optimized parameters like facet angle, size, and distance, along with cell radius edge ratio. However, the cell size (set to 3.0 in this case) had the most significant impact on determining the overall mesh coarseness or fineness. A smaller cell size results in a finer mesh with more elements and nodes, leading to a more detailed representation of the breast anatomy. The generated meshes exhibited variation in element and node count depending on the specific breast anatomy. The element count ranged from a minimum of 30,960 to a maximum of 157,745 , with an average of 96,375 elements (tetrahedrons). Similarly, the number of nodes varied between 6,196 and 28,450, with an average of 17,170. Figure 4 is illustrative of the FE volume mesh of the third phantom, which consists of 95,865 elements and 17,595 nodes. Material properties for the skin, glandular, and adipose components of the breast were assigned based on preclassified voxelized data. The assignment criteria were as follows:

1. Elements were labeled by mapping element centers to corresponding pixel values from the preclassified image.
2. Any pixel value of air was adjusted to skin properties, ensuring a continuous skin layer around the breast.
3. An additional label was assigned to elements near the chest wall, restricting motion in the anterior-posterior direction while allowing degrees of freedom in the superior-inferior and medial-lateral directions as depicted in Figure 5.

The meshes consisted of continuous solid elements without contact interfaces between different materials.

*3.2.2. Material model and boundary conditions*

The material model and boundary conditions define the breast's physical characteristics. Acquiring an accurate material model tailored to each patient's data is impractical. In this work, the same material model is applied to all patients, with consistent elastic constants assigned to each tissue type. An isotropic hyperelastic Neo-Hookean material model was employed to represent the nonlinear and incompressible behavior of the three tissues of the breast during deformation (Wellman et al., 1999). The material properties were assigned stiffness measures using Young's modulus ($E_{\text{adipose}}$ = 4.46 kPa, $E_{\text{glandular}}$ = 15.1 kPa, $E_{\text{skin}}$ = 20.0 kPa), while a nearly incompressible Poisson's ratio of 0.49 was set for all three tissues to describe the stress-strain relationship of the breast tissue.

Anatomically, the breast is not firmly attached to the body but rather rests on the thorax, connected by tissue that permits slight movement along the thorax. During mammographic imaging, the plates do not compress the breast adequately to induce displacement relative to the thorax. Therefore boundary conditions were established to restrict the rigid motion of the breast and simulate the connection between the breast and the body.

1. The outer triangular faces of the elements representing the chest wall surface were constrained in the anterior-posterior direction, see Figure 5.
2. The faces located superior or inferior to the axial midplane of the breast and in contact with plates were constrained in the medial-lateral direction.

The implementation of these boundary conditions permits unrestricted interaction between the breast mesh and the opposing surfaces representing the compression plates. This unrestricted interaction is essential for model stability and convergence, leading to the successful completion of the simulation.

*3.2.3. Breast compression simulation*

To mimic mammographic acquisition in a craniocaudal projection, the breast is compressed due to the vertical movement of a superior plate that pushes the breast against an inferior plate. The plates were implicitly defined as two infinite rigid wall constraints in direct contact with the breast. The rigid walls were enforced using an *Augmented Lagrangian* approach with a prescribed displacement, the convergence tolerance for the Lagrange multipliers was set to 0.01 with a penalty factor of 1000 to control the rate of convergence. In order to replicate the process, the upper plate was moved towards the bottom plate along the longitudinal axis. Since the breast typically rests on the inferior plate during mammography, a minor upward offset of 18 mm was applied by the bottom plate to flatten the model and support the breast, as shown in Figure 6.

Certain physiological and practical factors were considered to achieve appropriate compression. For example, recognizing that the breast is not rigidly attached to the body, movement of faces on the chest wall side surface was restricted in the anterior-posterior direction. Additionally, it was assumed that the natural shape of the breast, influenced by patient positioning on a prone table and gravity, mimicked the manual adjustments made by radiographers during mammography. Hence, the effect of gravity on the reference state was neglected (García et al., 2020). The friction coefficient between the plates and the breast is unknown. Therefore, the contact problem was solved using frictionless contact, which involved two rigid walls (both plates) and a deformable body (the breast).

Since the corresponding mammograms were not available. The amount of breast compression was de-

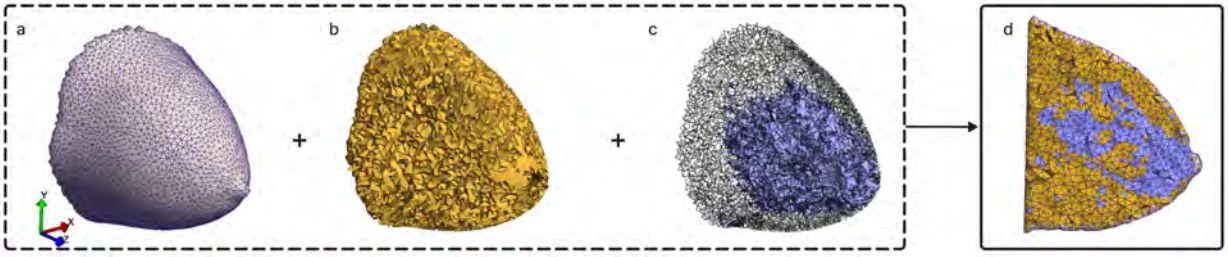



Figure 4: Finite element (FE) volume mesh model of phantom 3. (a) Shows the elements representing skin tissue. (b) Shows the elements corresponding to adipose tissue. (c) Shows the elements for glandular tissue. (d) Represents a cut view of the mesh, which consists of continuous solid elements without interfaces between different materials. The x-axis (red) indicates the mediolateral direction, the y-axis (green) represents the superior-inferior direction, and the z-axis (blue) represents the anterior-posterior direction

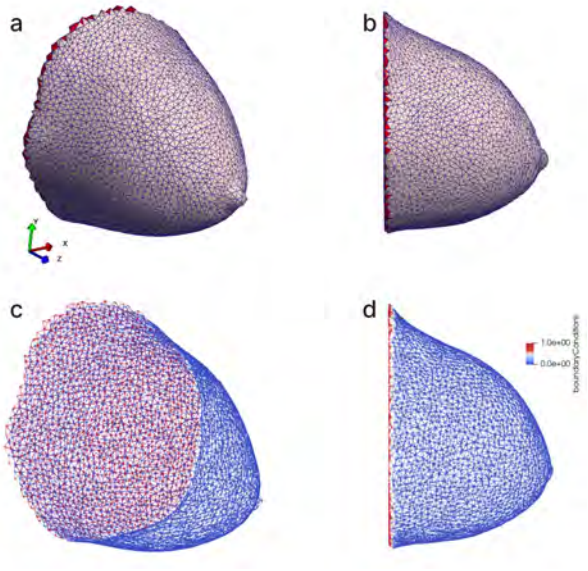



Figure 5: To illustrate the boundary conditions applied to the mesh surface, we can see the element faces color-coded to represent their boundary conditions. On the top: The outer faces of the tetrahedral elements on the posterior breast surface (red) are constrained in the anterior-posterior direction to mimic the chest wall restriction. The remaining elements throughout the breast are not restricted. On the bottom: Visualizing the mesh lines on the surface of the breast, Red lines represent faces with fixed anterior-posterior displacement (constrained), while blue lines represent faces free to move in all directions (unconstrained).

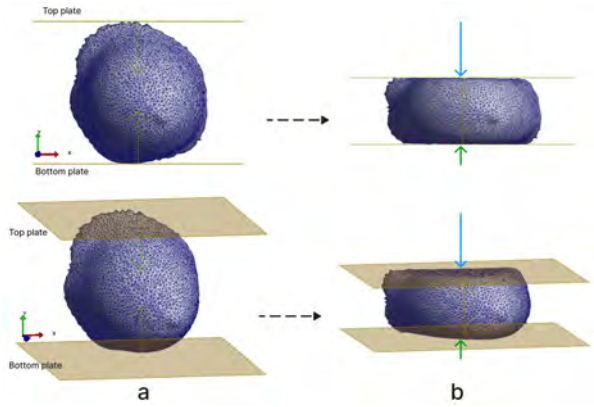



Figure 6: Compression simulation of a breast phantom using a FE model. The left images (a) show the undeformed breast mesh. The right images (b) depict the compressed breast mesh after simulating the compression process. During compression, the bottom plate (green arrows) was moved upwards by 18 mm to flatten the model. Simultaneously, the top plate (blue arrows) was moved downwards until the desired breast thickness was achieved.

rived from the thickness of the corresponding compressed phantoms.

### *3.2.4. Voxelised compressed phantom reconstruction*

The method for reconstructing the compressed phantom is based on the approach developed by (García et al., 2020). After compressing the biomechanical mesh model, it is converted into a voxelized breast phantom. A uniform grid around the model is used to store the elements. The grid's accuracy depends on internal factors (like grid resolution or voxel size) and external factors (such as the number of elements in the model or the degree of compression). The voxel size of 2 mm is used to define the grid, with the origin and size determined by the axis-aligned bounding box of the compressed model (García et al., 2017). Each voxel in the voxelized phantom is defined as a point within this grid.

For each point representing a voxel in the compressed breast phantom, the corresponding voxel in the grid is calculated from its physical position using barycentric coordinates. If the point lies within an element, these coordinates are used to locate the position in the uncompressed model. Hence, all points along a ray in the compressed model are transferred to the uncompressed

model, forming a curve in the BCT image. It's important to note that computing barycentric coordinates is a transformation from the world reference system $[x, y, z]$ to the model's internal reference system, denoted by $[E, \mathbf{b}] = [E, b_1, b_2, b_3, b_4]$, where $E$ is the element index and $\mathbf{b} = [b_1, b_2, b_3, b_4]$ are the barycentric coordinates. Each point in the physical space is uniquely represented by one vector $[E, \mathbf{b}]$ and vice versa (García et al., 2018).

The label corresponding to each point is directly acquired from the preclassified BCT volume through the method of nearest neighbor interpolation.

### *3.3. Deep learning for predicting breast deformation*

Efforts to enhance simulation speed have led to the development of deep-learning models aimed at replacing biomechanical simulations. Recent studies have suggested utilizing data-driven models generated by training different machine learning algorithms, such as random forests and artificial neural networks (ANNs), with FEA results. This approach aims to accelerate tissue deformation approximations through prediction. PhysGNN, a data-driven model has been developed to estimate the solution obtained by FEM using GNNs, which can incorporate mesh structural information and perform inductive learning on unstructured grids and complex topological structures (Salehi and Giannacopoulos, 2022). PhysGNN utilizes edge information by learning biomechanical deformation based on graphs created from FE meshes. In this work, we adapted PhysGNN for the breast compression application, as it demonstrated highly promising results.

#### *3.3.1. Architecture*

The depicted network architecture processes breast mesh data through a series of graph convolutional layers, specifically GraphConv and GraphSAGE layers, which use different aggregation methods (add and max). The architecture includes PReLU activation functions and Jumping Knowledge (JK) connections for feature combination and preservation. The spatial features are extracted in each GNN layer and aggregated using jumping-knowledge connections. After graph convolutions, the output is passed through linear layers with PReLU activations to transform the features, ultimately predicting displacements in the x, y, and z directions ($\delta x$, $\delta y$, $\delta z$). This design enables efficient and accurate modeling of mesh-based data for compression analysis. Figure 7 illustrates the architecture of the PhysGNN model used for predicting breast compression.

#### *3.3.2. Data generation and features construction*

The mesh model of Phantom 3 generated for the biomechanical model was used for generating the dataset. The PhysGNN method was designed to predict the gradual deformation of a 3D FE mesh in response to a series of planned incremental displacements. The ground truth was the nodal displacements obtained from NiftySim incremental simulations.

A basic approach to incremental simulations using deep learning entails breaking down a large simulation into a series of smaller, self-contained steps. These steps are then executed sequentially, essentially treating them as individual simulations. FEM simulations were divided into a series of incremental steps, to address the non-linear behavior of soft tissue. This approach simulates the material's response through multiple transitional states, providing a more accurate representation of the deformation process. To assess the effectiveness of PhysGNN in predicting breast compression imposed by force during mammography (modeled by compressing the breast between two plates in our work), we utilized several data points within the PhysGNN model. The material assignment to the mesh elements based on the preclassified image was used as material ID, allowing the model to distinguish between different tissue types. Additionally, the fixed nodes that belong to the chest wall, which are constrained, were used as boundary condition IDs. Finally, Young's modulus of the tissues was incorporated as a physical property within the PhysGNN model, providing information about the material's stiffness and influencing the predicted deformations. The surface nodes in the dataset are the nodes located on the outer surface of the breast mesh (1129). The dataset was created by applying a force of 90 Newtons to the breast surface at a time in 30-time steps and 40 directions to capture the non-linear behavior of soft tissue under large forces. Therefore, the amount of force applied to one of the surface nodes at time $i$ is:

$$\mathbf{F}_i = \frac{\mathbf{F}_{\text{total}}}{30 \times 1129} \times i, \quad i \in \{1, \ldots, 30\} \tag{1}$$

The forces applied to each surface node are directed along its surface normal ($x$, $y$, and $z$) and three additional directions, with each direction, represented as a tuple ($x$, $y$, and $z$) randomly sampled from a unit-radius hemisphere. The implementation of (Salehi and Giannacopoulos, 2022) included 10 distinct batches of random directions, along with the surface normal direction.

The features inputted into PhysGNN include the force values applied to surface nodes in both Cartesian coordinates $(F_x, F_y, F_z)$ and spherical coordinates $(F_\rho, F_\theta, F_\phi)$. Additionally, each node is assigned a constant value called Physical Property, which varies based on the node's boundary condition and tissue type. For nodes with a free boundary condition, the value is set to 0.1 for skin tissue, 0.6 for glandular tissue, and 1 for fat tissue. For nodes with a fixed boundary condition, the value is 0. This value determines the extent of displacement a node can undergo, influenced by its boundary condition and Young's modulus. Fat tissue, with its lower Young's modulus, can undergo larger displacements compared to glandular or skin tissues. Specifically, the value 0.1 is derived from the ratio of fat

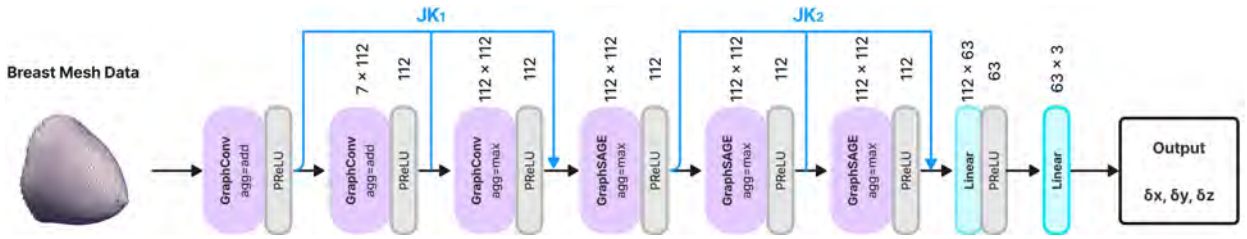



Figure 7: Architectural diagram of PhysGNN (Salehi and Giannacopoulos, 2022).

tissue's Young's modulus (4.46 KPa) to the combined Young's modulus of gland and skin tissues (15.1 KPa + 20.0 KPa). Similarly, the value 0.6 is calculated from the ratio of gland tissue's Young's modulus (15.1 KPa) to the combined Young's modulus of fat and skin tissues (4.46 KPa + 20.0 KPa). PhysGNN outputs the displacement values of the mesh nodes in the $x$, $y$, and $z$ directions ($\delta_x, \delta_y, \delta_z$). For the GNN models, edge weights ($e_{u,v}$) are computed as the inverse of the Euclidean distance between adjacent nodes $u$ and $v$, with $u, v \in V$ as:

$$e_{u,v} = \frac{1}{\sqrt{(x_u - x_v)^2 + (y_u - y_v)^2 + (z_u - z_v)^2}} \tag{2}$$

### 3.3.3. Hyperparameters of PhysGNN

Similar to prior research (Martínez-Martínez et al., 2017; Said et al., 2023), the predictions generated by PhysGNN data-driven model, are evaluated against the results obtained from FEM, which serves as the ground truth.

The loss function used for learning the trainable parameters is the mean Euclidean error computed as:

$$MEE = \frac{1}{\mathcal{N}} \sum_{n \in \mathcal{N}} \sqrt{\sum_{i=1}^{3} \left(y_n^i - \hat{y}_n^i\right)^2} \tag{3}$$

where $\mathcal{N}$ is the number of mesh nodes, $\mathbf{y} \in \mathbb{R}^{\mathcal{N}\times 3}$ represents the FEM-approximated displacement in the $x$, $y$, and $z$ directions, and $\mathbf{z} \in \mathbb{R}^{\mathcal{N}\times 3}$ represents the displacement predicted by PhysGNN. The AdamW optimizer, with an initial learning rate of 0.005, was used to minimize the loss value, reducing the rate by a factor of 0.1 to a minimum of $1 \times 10^{-8}$ if validation loss did not improve after 5 epochs. Early stopping, halting training after 15 epochs without validation loss improvement, was employed to prevent overfitting. Additionally, a dropout rate of 0.1 was applied to the penultimate layer of PhysGNN to enhance generalization. The model was trained in 8 batches for faster convergence.

Table 2 summarizes the input features (X) and corresponding output values (Z) processed by the PhysGNN model, an additional information on data generation is in Appendix A.

Table 2: PhysGNN features and outputs.

| Features ($X$) | Output ($Z$) |
|---|---|
| $F_x, F_y, F_z, F_\rho, F_\theta, F_\phi, PhysicalProperty$ | $\delta x, \delta y, \delta z$ |

### 3.3.4. PhysGNN training experiments

To train and evaluate the performance of our models, we divided the dataset into three subsets: training, validation, and testing. The training set provided the model with examples to learn the underlying relationships between the input data (representing breast geometry) and the desired output (deformation patterns). The validation set played a crucial role, as it was used to fine-tune the models' hyperparameters without directly influencing their performance on unseen data.

To evaluate the models' performance, we employed two contrasting data partitioning strategies: leave-one-deformation-out and hold-out. Leave-one-deformation-out, a geometry-based approach, prioritizes keeping entire breast configurations (representing different deformation states) together. During partitioning, for each deformation state, all data points are used for training and validation purposes except for one. This single data point, representing a specific deformation state of a particular breast, is isolated and reserved for the test set. In our case, we choose the final compression state for the test set. This ensures the models are exposed to a wide range of completely deformed breast shapes during training while offering unseen deformations for testing their generalizability. In contrast, hold-out focuses on individual data points, disregarding the geometric context (i.e., the relationship between different deformation states of the same breast). This instance-based approach can scatter data points from a single deformation sequence across different sets (training, validation, or testing). By employing these contrasting strategies, we aimed to gain a comprehensive understanding of how the partitioning strategy affects the performance of the model.

1. *Hold-out* experiment involved randomly splitting the generated dataset into training (70%), validation (20%), and testing (10%) sets. As mentioned earlier, this approach disregards the breast deformation state during data partitioning. Despite having a single breast phantom with 30 deformation

states, we conducted a hold-out experiment to assess how well the model generalized to unseen deformations within the same breast. While acknowledging the limitations of using a single breast for training and validation, this experiment offers valuable groundwork for future studies with additional breast models.

2. *Leave-one-deformation-out* experiment assessed the models' ability to generalize to unseen deformations, we adopted a unique testing strategy. For the 30 deformations simulated in NiftySim, a single deformation was isolated for testing, while the remaining 29 deformations were split into training (80%) and validation (20%) sets. This approach mimics real-world clinical scenarios where models are trained on a collection of known deformations and then tasked with predicting the behavior of a new, unseen deformation for a complete breast geometry.

### *3.4. Qualitative and Quantitative Results of Finite Element Analysis*

A total of 60 phantoms were included in this investigation for FEA. However, convergence difficulties limited the number of successfully analyzed phantoms to 35. The remaining 25 phantoms exhibited convergence issues during the FE simulations. It is hypothesized that the presence of additional axillary tissue in the segmentation, which is not anatomically part of the breast, might have hindered the convergence process in these cases. This issue needs further investigation in future work.

Figure 8 illustrates a sample of voxelized phantoms generated from the previously compressed volumes reported in (Sarno et al., 2021b), as well as those simulated using NiftySim and FEBio.

BCT phantoms revealed a pre-compression breast thickness ranging from 85 mm to 169 mm (mean: 132 mm ± 21 mm). After compression, the expected thickness falls within a range of 42 mm to 89 mm (mean: 63 mm ± 12 mm). This translates to an average compression ratio of 0.48 ± 0.12.

It is evident that the FEM-based results (NiftySim and FEBio) are more realistic and reliable, closely mimicking real-world applications. Both NiftySim and FEBio produce very similar outcomes, indicating their consistency in modeling. Closer inspection reveals that NiftySim potentially delivers a more precise compression thickness compared to the compressed phantoms from the dataset. The primary distinctions in compression for FEBio are observed in the mediolateral and anterior directions when compared to the other methods.

The total breast phantom volume ($mm^3$), glandular tissue volume ($mm^3$), and volumetric breast density (VBD in %) were calculated for both the uncompressed and compressed phantoms.

NiftySim delivered accurate approximations of the compressed phantom thickness, highlighting its effectiveness in simulating tissue deformation. Moreover, by leveraging GPU acceleration, NiftySim achieved efficient execution times, averaging 300 seconds per simulation with a standard deviation of 60 seconds. The VBD was calculated as the total glandular tissue divided by the total breast volume. VBD ranged between 0.86 % and 24.97 % (mean: 9.63 %) before compression and between 0.87 % to 25.30 % (mean: 9.75 %) after compression. The mean difference between the uncompressed and the compressed VBD was 0.12%, showing a correlation coefficient (R) of 0.99. Figure 9(a) shows the correlation of breast volume, glandular tissue volume, and VBD, before and after the compression process.

FEBio approximations were close to the expected thicknesses. FEBio simulations on CPU hardware required an average of 6634 seconds to complete, with a standard deviation of 4919 seconds. VBD ranged between 0.86 % and 24.97 % (mean: 9.74 %) before compression and between 0.87 % to 25.37 % (mean: 9.88 %) after compression. The mean difference between the uncompressed and the compressed VBD was 0.14%, showing a correlation coefficient (R) of 0.98. Figure 9(b) shows the correlation of breast volume, glandular tissue volume, and VBD, before and after the compression process.

Approximating tissue deformation using FEBio was carried out on an Intel(R) Xeon(R) Silver 4208 CPU with 128 GB.

Table 3 summarises the mean values for Total breast volume loss (%), fat tissue volume loss (%), glandular tissue volume loss (%), and skin tissue volume loss (%). FEBio simulations seem to show a greater loss of overall breast tissue, including fat and glandular tissue, during compression compared to NiftySim. Conversely, NiftySim simulations appear to predict a higher loss of skin tissue during compression.

The DICE score is a common metric used to quantify the spatial overlap between two images. Higher DICE scores indicate greater similarity. Table 4 shows the mean and standard deviation of the DICE score, which compares the similarity between the reconstructed compressed phantoms obtained from NiftySim and FEBio, for all 35 phantoms included in the study. Several factors could contribute to the observed variations in Dice scores across tissue types; such as the mesh generation process can influence how tissues are represented. Also, the specific algorithms used by NiftySim and FEBio to simulate the compression process might differ. This could lead to variations in how each tissue type responds to compression. This warrants further investigation in future work.

Figure 10 presents a comparison of displacement magnitudes obtained from two FEA solvers: NiftySim and FEBio. (a) displays the output displacements pre-

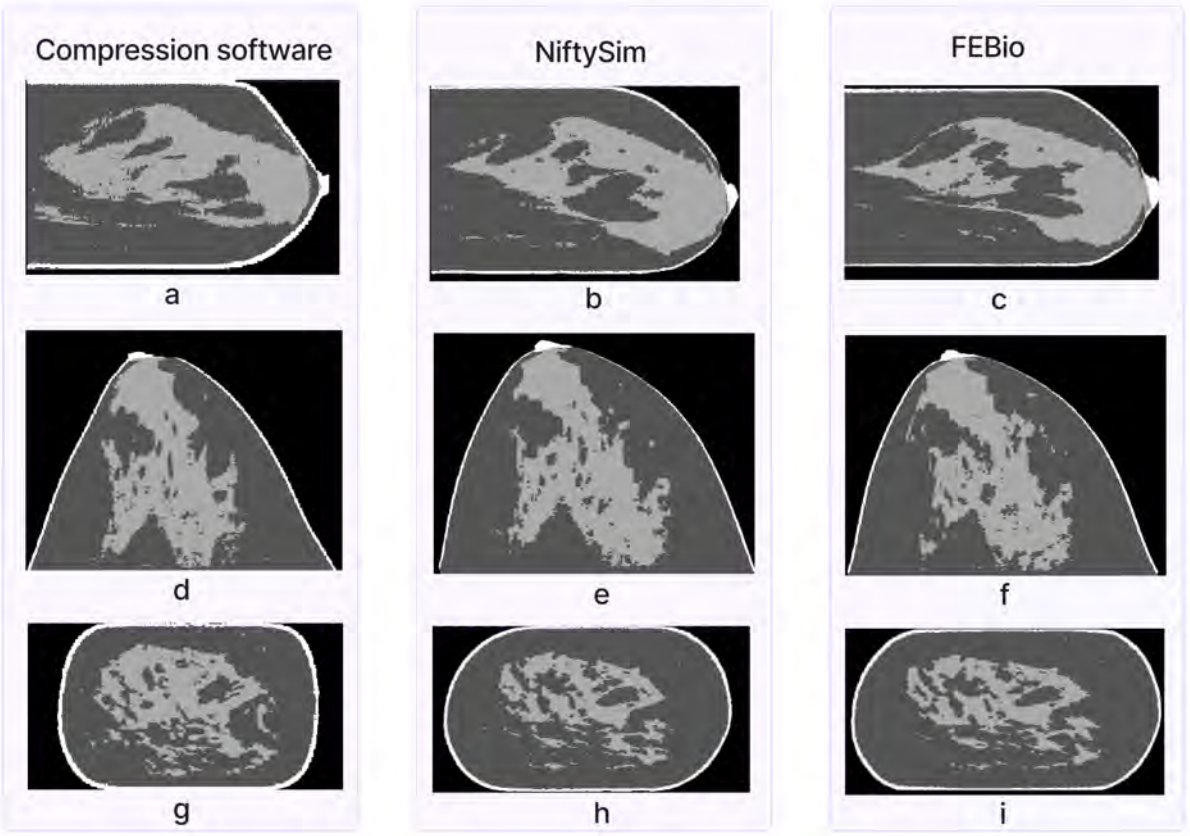



Figure 8: Cross-sectional of compressed digital phantoms from the compression software in (Zyganitidis et al., 2007), NiftySim and FEBio. (a), (b), and (c) are for sagittal cross sections. (d), (e), and (f) are for axial cross sections. (g), (h), and (i) are for coronal cross sections.

Table 3: Breast tissue loss during compression

| FEM | Total breast volume loss(%) | Fat tissue volume loss(%) | Glandular tissue volume loss(%) | Skin tissue volume loss(%) |
|---|---|---|---|---|
| NiftySim | 1.18 ± 0.34 | 0.34 ± 0.16 | 0.09 ± 0.11 | 12.41 ± 2.97 |
| FEBio | 1.70 ± 0.70 | 0.99 ± 0.76 | 0.42 ± 0.45 | 11.97 ± 2.32 |

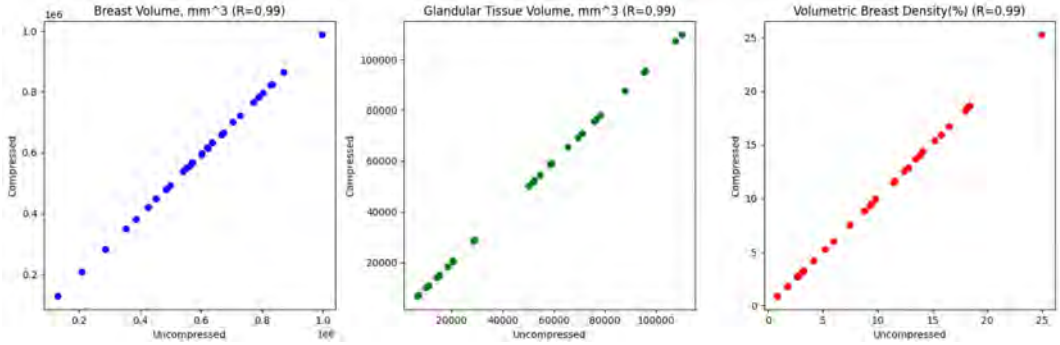
Table 4: Mean Dice score between all the 35 reconstructed compressed phantoms obtained from NiftySim and FEBio

| Fat Tissue *Mean Dice* | Glandular Tissue *Mean Dice* | Skin Tissue *Mean Dice* |
|---|---|---|
| 0.78 ± 0.16 | 0.38 ± 0.221 | 0.16 ± 0.15 |

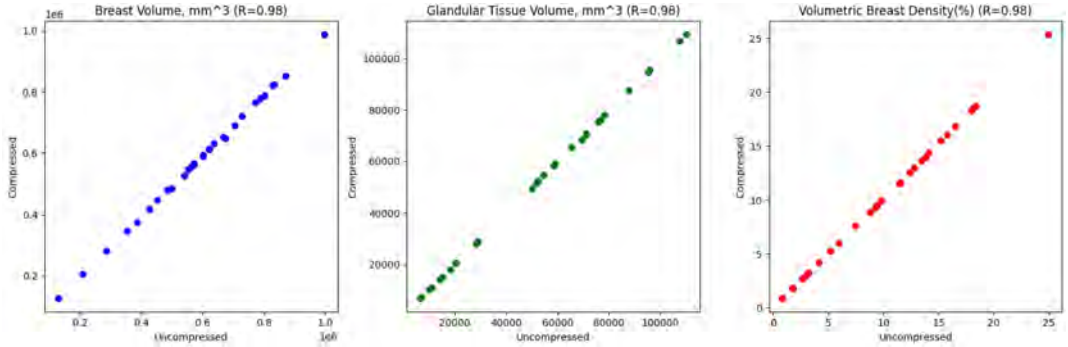
dicted by NiftySim for the FE model. (b) shows the corresponding output displacements from FEBio. Finally, (c) depicts the difference between the displacement magnitudes calculated by NiftySim and FEBio, highlighting the discrepancies in their FEA approximations. While the overall displacement patterns appear similar, there are noticeable differences in the top and outer regions. These variations could be due to compression plates or the specific boundary conditions applied to the biomechanical model.

### *3.5. Performance of PhysGNN model*

To train the deep learning model, a dataset was generated from the results of incremental FE simulations performed on only one successfully analyzed phantom (phantom 3).




(a) The compression was obtained by NifySim.




(b) The compression was obtained using FEBio.

Figure 9: Correlation of breast volume, glandular tissue volume, and VBD before and after the compression. 35 BCT phantoms were analyzed from NiftySim and FEBio. Correlation coefficients of 0.99, and 0.98 were observed, respectively.

According to the results in Table 5, PhysGNN is capable of effectively predicting tissue deformation under prescribed force loads, especially highlighted by 99.96% and 81.22% of absolute position errors being under 1 mm in hold-out and leave-one-deformation-out experiments, respectively.

The prediction of tissue deformation per each FE simulation on an Intel(R) Xeon(R) E5-2630 v4 @ 2.20GHz CPU took 0.4161 ± 0.0426 seconds on average, while on an NVIDIA GeForce RTX 2080 Ti GPU with 46 GB took 0.01 ± 0.06 seconds on average.

The hold-out yielded better and more stable predictive performance estimates than leave-one-deformation-out due to lower variance, and computational efficiency. While, the latter is computationally expensive, limiting model complexity and practical hyperparameter tuning, and single-point testing may not reflect realistic performance on diverse datasets.

Table 6 reflects the test set statistics of the PhysGNN experiments.

In the hold-out experiment, predicting tissue deformation took 0.42 ± 0.04 seconds on CPU and 0.01 ± 0.06 seconds on GPU. In the leave-one-deformation-out experiment, it took 0.82 seconds on CPU and 0.47 seconds on GPU. Incremental NiftySim simulations, with GPU acceleration, totaled 4640.5 seconds (154.7 seconds per simulation). This indicates a speedup of 329 times with GPU and 188 times with CPU in the leave-one-deformation-out experiment compared to a single NiftySim simulation.

### *3.6. Qualitative and Quantitative Results of PhysGNN in Leave-one-deformation-out experiment*

We present here additional quantitative and qualitative results only for the leave-one-deformation-out experiment. For the hold-out experiment, since nodes from different deformations are mixed in the test set, a specific evaluation cannot be provided.

The reconstructed phantom of predicted deformation by PhysGNN in the leave-one-deformation-out experiment compared to the reconstructed phantom of the NiftySim displacement (Ground Truth) is shown in Figure 11, indicating that the PhysGNN approximation of the breast compression is very similar to the approximation achieved by FEM NiftySim. This close similarity highlights the effectiveness of PhysGNN in accurately modeling and predicting breast tissue deformation, validating its potential as a reliable alternative to traditional FEM methods. The similarity is further quantified by the Dice scores in Table 7, showing high values for pri-

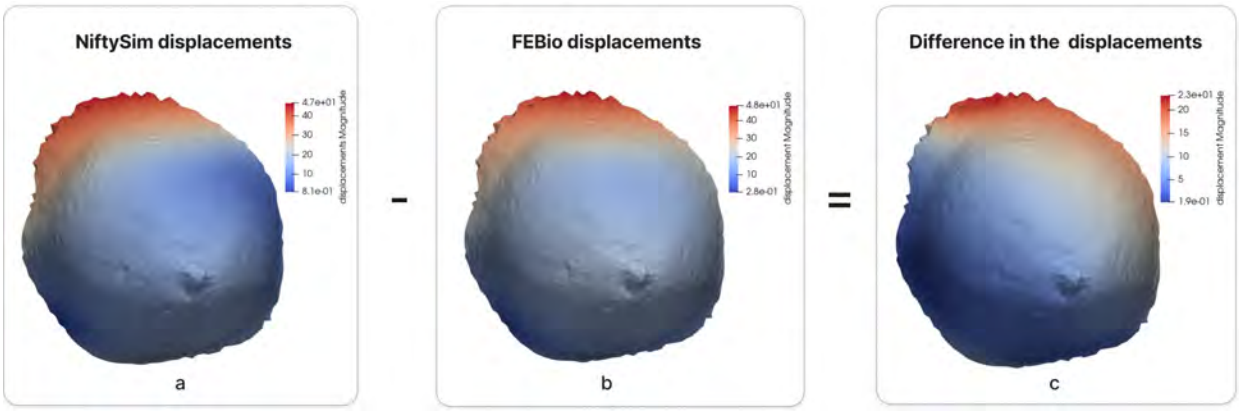



Figure 10: Output displacements in mm of the FEM for phantom 3.

Table 5: The performance of PhysGNN model on Hold-out and Leave-one-deformation-out of the test set.

| Experiment | MAE ($\delta x$) (mm) | MAE ($\delta y$) (mm) | MAE ($\delta z$) (mm) | Mean Euclidean Error (mm) | Euclidean Error ≤ 1 mm (%) | Mean Absolute Position Error (mm) | Absolute Position Error ≤ 1 mm (%) |
|---|---|---|---|---|---|---|---|
| Hold-out | 0.17 ± 0.18 | 0.20 ± 0.20 | 0.13 ± 0.14 | 0.34 ± 0.15 | 97.50 | 0.17 ± 0.03 | 99.96 |
| Leave-one-deformation-out | 0.56 ± 0.44 | 0.52 ± 0.39 | 0.67 ± 0.52 | 1.21 ± 0.44 | 33.51 | 0.58 ± 0.06 | 81.22 |

Table 6: The test set statistics of Hold-out and Leave-one-deformation-out, where $y$ is the displacement, and Max. Euclidean $\text{Error}_{\text{mean}}$ is computed as the average of the maximum Euclidean error observed for each data element—i.e., each simulation.

| Experiment | $\delta y_{\text{max}}$ (mm) | $\delta y_{\text{mean}}$ (mm) | Max. Euclidean $\text{Error}_{\text{mean}}$ (mm) |
|---|---|---|---|
| Hold-out | 46.55 | 24.02 ± 13.26 | 2.60 ± 1.40 |
| Leave-one-deformation-out | 46.55 | 46.55 ± 0.00 | 5.16 ± 0.00 |

mary tissues (excluding skin, which is very thin and thus results in a compromised Dice score).

Finally, Table 8 summarizes the volume loss percentages for total breast, fat, glandular, and skin tissues. PhysGNN predictions show a slightly increased overall breast tissue loss.

Table 7: Dice score of reconstructed compressed phantoms obtained from NiftySim and PhysGNN

| Fat Tissue | Glandular Tissue | Skin Tissue |
|---|---|---|
| 0.94 | 0.83 | 0.53 |

We can also visually assess the displacement magnitudes on the BCT model with each method as shown in Figure 12. Figure 12 (a) shows NiftySim's displacements, Figure 12 (b) shows PhysGNN's predicted displacements, and Figure 12 (c) highlights the differences. While overall patterns are similar, noticeable dissimilarity exists on the breast surface, with a maximum displacement difference of 5.2 mm.

## 4. Discussion

The FEM provides a more realistic compression of the breast phantoms compared to the compression software (Zyganitidis et al., 2007) used in generating the

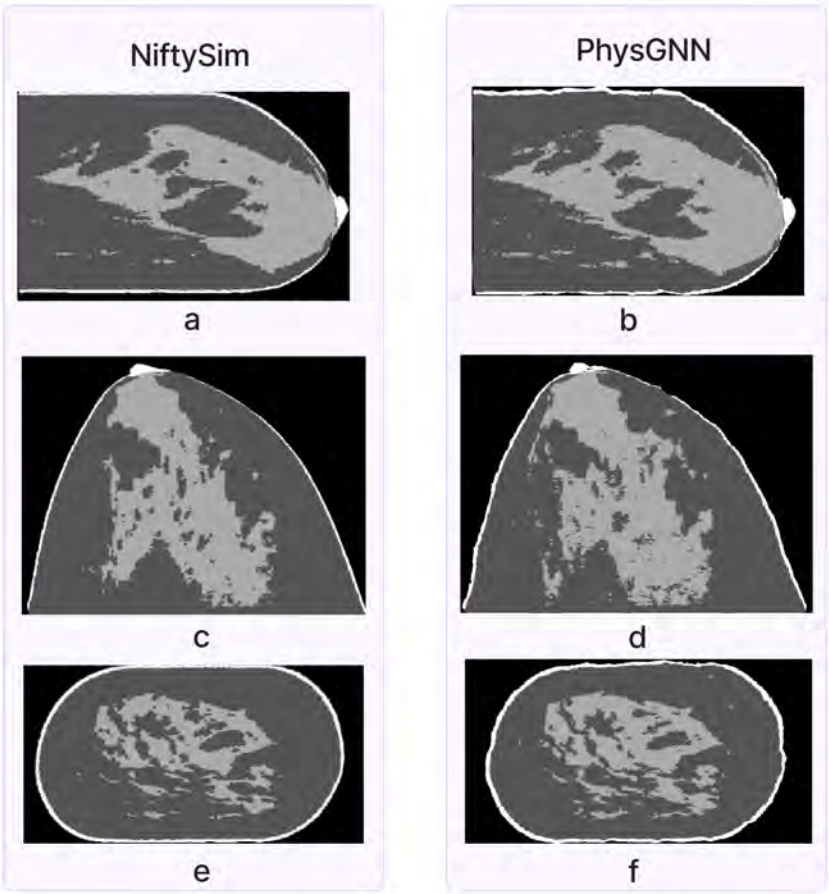



Figure 11: Cross-sectional of compressed digital phantoms from NiftySim and PhysGNN. (a), (b) are for sagittal cross sections. (c), (d) are for axial cross sections. (e), (f) are for coronal cross-sections.

Table 8: Breast tissue loss during compression

| FEA/DL | Total breast volume loss(%) | Fatty tissue volume loss(%) | Glandular tissue volume loss(%) | Skin tissue volume loss(%) |
|---|---|---|---|---|
| NiftySim | 1.03 | 0.38 | 0.03 | 10.33 |
| PhysGNN | 1.26 | 0.55 | 0.34 | 10.94 |

compressed phantoms (Sarno et al., 2021b). The FEM methods produce nearly identical simulation outputs, displacements, and material behavior under compression. Various factors can influence the accuracy of FEM approximations such as; mesh density and element quality, with finer meshes typically yielding more accurate results. The observed differences may arise from variations in the model formulation, boundary condition definition, and the specific analysis techniques employed by each FEM.

PhysGNN manifests itself as a promising deep learning module for predicting tissue deformation due to its accurate approximations, computation speed, and ease of implementation. A limitation of the model is that it is trained on several states of one mesh geometry. Incorporating more geometries means the meshes have to have the same number of elements and nodes. Further investigation on applying PhysGNN to different meshes is planned for future work. Another limitation of PhysGNN is that it requires incremental FEM simulations to be trained. While incremental FEM simulations are very accurate, they require significant computational time, sometimes approaching 5 minutes (in the case of NiftySim) to perform a single simulation. This greatly increases training requirements. On the other hand, once PhysGNN is trained, the inference simulation time is considerably faster than FEM, which can take several minutes to hours to perform an incremental

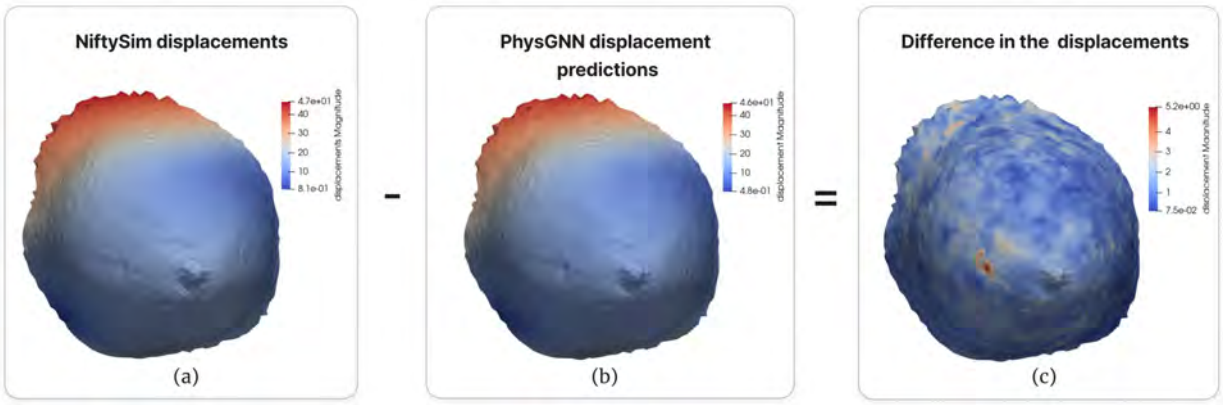



Figure 12: Output displacements (mm) of NiftySim and PhysGNN on the FE model

simulation. In our experimentation to optimize prediction accuracy, we explored various modifications to the model's parameters and architecture. Specifically, we tested increasing the number of random directions input and the number of GNN layers. However, these modifications did not yield the expected improvements. We attempted to increase the number of random directions to seven; however, the prediction results were not as accurate as those obtained with four directions, leading us to discontinue further investigation in this area. Additionally, we explored increasing the number of GNN layers to nine, but similar to the previous case, this did not enhance the predictions and proved to be computationally expensive, resulting in longer processing times.

## 5. Conclusions

This thesis has presented a comparison and evaluation of Finite Element Methods (FEM) for building breast models for breast compression simulation. Building on the earlier work of García, in this thesis we have performed a larger evaluation using BCT data, incorporated and compared additional FE solvers such as FEBio, and investigated the effects of mesh parameters and boundary conditions. Moreover, this thesis has presented the first results on the applicability of a deep learning-based approach (PhysGNN) for simulating breast deformation with quantitative and qualitative comparison to the standard FEM models, showing its accuracy and potential applicability in these scenarios. Notably, The PhysGNN model, trained on FEM displacement data, achieved a Mean Euclidean Error of $0.34 \pm 0.15$ mm in the hold-out experiment.

## Acknowledgments

I would like to express my deepest gratitude to my supervisor, Professor Robert Martí Marly, for his invaluable guidance, support, and encouragement throughout my research. His expertise and insightful feedback have been instrumental in the completion of this thesis.

I would also like to extend my heartfelt thanks to my co-supervisor, Dr. Eloy García Marcos, for his continuous support and assistance. His constructive suggestions and dedicated mentorship have significantly contributed to the progress and success of my work.

Additionally, I would like to thank the entire faculty and staff of ViCOROB, for providing a conducive environment for my studies and research. I am also grateful to my family and friends for their unwavering support and encouragement.

## References

Azar, F., Metaxas, D., Schnall, M., 2000. A finite element model of the breast for predicting mechanical deformations during biopsy procedures, in: Proceedings IEEE Workshop on Mathematical Methods in Biomedical Image Analysis. MMBIA-2000 (Cat. No.PR00737), pp. 38–45. doi:10.1109/MMBIA.2000.852358.

Bakic, P.R., Zhang, C., Maidment, A.D., 2011. Development and characterization of an anthropomorphic breast software phantom based upon region-growing algorithm. Medical physics 38, 3165–3176.

Boone, J.M., Nelson, T.R., Lindfors, K.K., Seibert, J.A., 2001. Dedicated breast ct: Radiation dose and image quality evaluation. Radiology 221, 657–667. doi:10.1148/radiol.2213010334. pMID: 11719660.

Chung, J., Rajagopal, V., Nielsen, P., Nash, M., 2008. A biomechanical model of mammographic compressions. Biomechanics and Modeling in Mechanobiology 7, 43–52. doi:10.1007/S10237-006-0074-6.

García, E., Diez, Y., Diaz, O., Lladó, X., Martí, R., Martí, J., Oliver, A., 2018. A step-by-step review on patient-specific biomechanical finite element models for breast mri to x-ray mammography registration. Medical physics 45, e6–e31.

García, E., Diez, Y., Diaz, O., Lladó, X., Gubern-Mérida, A., Martí, R., Martí, J., Oliver, A., 2018. Multimodal breast parenchymal patterns correlation using a patient-specific biomechanical model. IEEE Transactions on Medical Imaging 37, 712–723.

García, E., Diez, Y., Diaz, O., Lladó, X., Gubern-Mérida, A., Martí, R., Martí, J., Oliver, A., 2019. Breast mri and x-ray mammography registration using gradient values. Medical Image Analysis 54, 76–87. doi:https://doi.org/10.1016/j.media.2019.02.013.

García, E., Fedon, C., Caballo, M., Martí, R., Sechopoulos, I., Diaz, O., 2020. Realistic compressed breast phantoms for medical physics applications, p. 73. doi:10.1117/12.2564273.

García, E., Oliver, A., Diaz, O., Diez, Y., Gubern-Mérida, A., Martí, R., Martí, J., 2017. Mapping 3d breast lesions from full-field digital mammograms using subject-specific finite element models, p. 1013504. doi:10.1117/12.2255957.

Han, L., Hipwell, J.H., Eiben, B., Barratt, D., Modat, M., Ourselin, S., Hawkes, D.J., 2013. A nonlinear biomechanical model based registration method for aligning prone and supine mr breast images. IEEE transactions on medical imaging 33, 682–694.

Han, L., Hipwell, J.H., Tanner, C., Taylor, Z., Mertzanidou, T., Cardoso, J., Ourselin, S., Hawkes, D.J., 2011. Development of patient-specific biomechanical models for predicting large breast deformation. Physics in Medicine Biology 57, 455. URL: https://dx.doi.org/10.1088/0031-9155/57/2/455, doi:10.1088/0031-9155/57/2/455.

Hipwell, J.H., Vavourakis, V., Han, L., Mertzanidou, T., Eiben, B., Hawkes, D.J., 2016. A review of biomechanically informed breast image registration. Physics in Medicine & Biology 61, R1.

Hopp, T., Baltzer, P., Dietzel, M., Kaiser, W.A., Ruiter, N.V., 2012. 2d/3d image fusion of x-ray mammograms with breast mri: visualizing dynamic contrast enhancement in mammograms. International journal of computer assisted radiology and surgery 7, 339–348.

Hopp, T., Dietzel, M., Baltzer, P., Kreisel, P., Kaiser, W., Gemmeke, H., Ruiter, N., 2013. Automatic multimodal 2d/3d breast image registration using biomechanical fem models and intensity-based optimization. Medical Image Analysis 17, 209–218. doi:https://doi.org/10.1016/j.media.2012.10.003.

Hsu, C.M., Palmeri, M.L., Segars, W.P., Veress, A.I., Dobbins III, J.T., 2013. Generation of a suite of 3d computer-generated breast phantoms from a limited set of human subject data. Medical physics 40, 043703.

Hsu, C.M.L., Palmeri, M.L., Segars, W.P., Veress, A.I., Dobbins III, J.T., 2011. An analysis of the mechanical parameters used for finite element compression of a high-resolution 3d breast phantom. Medical Physics 38, 5756–5770. doi:https://doi.org/10.1118/1.3637500.

Johnsen, S.F., Taylor, Z.A., Clarkson, M.J., et al., 2015. Niftysim: A gpu-based nonlinear finite element package for simulation of soft tissue biomechanics. International Journal of Computer Assisted Radiology and Surgery (Int J CARS) 10, 1077–1095. doi:10.1007/s11548-014-1118-5.

Lee, A.W., Rajagopal, V., Babarenda Gamage, T.P., Doyle, A.J., Nielsen, P.M., Nash, M.P., 2013. Breast lesion co-localisation between x-ray and mr images using finite element modelling. Medical Image Analysis 17, 1256–1264. doi:https://doi.org/10.1016/j.media.2013.05.011.

Li, C.M., Segars, W.P., Tourassi, G.D., Boone, J.M., Dobbins III, J.T., 2009. Methodology for generating a 3d computerized breast phantom from empirical data. Medical physics 36, 3122–3131.

Liu, Y.L., Liu, P.Y., Huang, M.L., Hsu, J.T., Han, R.P., Wu, J., 2017. Simulation of breast compression in mammography using finite element analysis: A preliminary study. Radiation Physics and Chemistry 140, 295–299. doi:https://doi.org/10.1016/j.radphyschem.2017.01.017. 2nd International Conference on Dosimetry and its Applications (ICDA-2) University of Surrey, Guildford, United Kingdom, 3-8 July 2016.

Maas, S.A., Ellis, B.J., Ateshian, G.A., Weiss, J.A., 2012. FEBio: Finite Elements for Biomechanics. Journal of Biomechanical Engineering 134, 011005.

Martínez-Martínez, F., Rupérez-Moreno, M., Martínez-Sober, M., Solves-Llorens, J., Lorente, D., Serrano-López, A., Martínez-Sanchis, S., Monserrat, C., Martín-Guerrero, J., 2017. A finite element-based machine learning approach for modeling the mechanical behavior of the breast tissues under compression in real-time. Computers in Biology and Medicine 90, 116–124. doi:https://doi.org/10.1016/j.compbiomed.2017.09.019.

Mendizabal, A., Tagliabue, E., Brunet, J.N., Dall'Alba, D., Fiorini, P., Cotin, S., 2020. Physics-based deep neural network for real-time lesion tracking in ultrasound-guided breast biopsy, in: Computational Biomechanics for Medicine: Solid and Fluid Mechanics for the Benefit of Patients 22, Springer. pp. 33–45.

Mertzanidou, T., Hipwell, J., Johnsen, S., Han, L., Eiben, B., Taylor, Z., Ourselin, S., Huisman, H., Mann, R., Bick, U., Karssemeijer, N., Hawkes, D., 2014. Mri to x-ray mammography intensity-based registration with simultaneous optimisation of pose and biomechanical transformation parameters. Medical Image Analysis 18, 674–683. doi:https://doi.org/10.1016/j.media.2014.03.003.

Mettivier, G., Sarno, A., Boone, J.M., Bliznakova, K., di Franco, F., Russo, P., 2020. Virtual clinical trials in 3D and 2D breast imaging with digital phantoms derived from clinical breast CT scans, in: Chen, G.H., Bosmans, H. (Eds.), Medical Imaging 2020: Physics of Medical Imaging, International Society for Optics and Photonics. SPIE. p. 1131259. doi:10.1117/12.2548224.

Miller, K., Joldes, G., Lance, D., Wittek, A., 2007. Total lagrangian explicit dynamics finite element algorithm for computing soft tissue deformation. Communications in numerical methods in engineering 23, 121–134.

Moreira, P., Peterlik, I., Herink, M., Duriez, C., Cotin, S., Misra, S., 2013. Modelling prostate deformation: Sofa versus experiments. Prostate 17, 1–0.

Pasciak, J.E., 1995. The mathematical theory of finite element methods (susanne c. brenner and l. ridgway scott). Siam Review 37, 472–473.

Pfaff, T., Fortunato, M., Sanchez-Gonzalez, A., Battaglia, P.W., 2020. Learning mesh-based simulation with graph networks. arXiv preprint arXiv:2010.03409 .

Phellan, R., Hachem, B., Clin, J., Mac-Thiong, J.M., Duong, L., 2021a. Real-time biomechanics using the finite element method and machine learning: Review and perspective. Medical Physics 48, 7–18.

Phellan, R., Hachem, B., Clin, J., Mac-Thiong, J.M., Duong, L., 2021b. Real-time biomechanics using the finite element method and machine learning: Review and perspective. Medical Physics 48, 7–18.

Ruiter, N., Müller, T., Stotzka, R., Gemmeke, H., Reichenbach, J., Kaiser, W., 2002. Automatic image matching for breast cancer diagnostics by a 3d deformation model of the mamma. Biomedical Engineering / Biomedizinische Technik 47, 644–647. doi:10.1515/bmte.2002.47.s1b.644.

Ruiter, N., Stotzka, R., Muller, T.O., Gemmeke, H., Reichenbach, J., Kaiser, W., 2006. Model-based registration of x-ray mammograms and mr images of the female breast. IEEE Transactions on Nuclear Science 53, 204–211. doi:10.1109/TNS.2005.862983.

Rupérez, M., Martínez-Martínez, F., Martínez-Sober, M., Lago, M., Lorente, D., Bakic, P., Serrano-López, A., Martínez-Sanchis, S., Monserrat, C., Martín-Guerrero, J., 2018. Modeling the mechanical behavior of the breast tissues under compression in real time, in: VipIMAGE 2017: Proceedings of the VI ECCOMAS Thematic Conference on Computational Vision and Medical Image Processing Porto, Portugal, October 18-20, 2017, Springer. pp. 583–592.

Said, S., Yang, Z., Clauser, P., Ruiter, N., Baltzer, P., Hopp, T., 2023. Estimation of the biomechanical mammographic deformation of the breast using machine learning models 11orcid (s): https://orcid.org/0000-0002-7441-6135 (s. said); https://orcid.org/0000-0002-4411-4056 (n.v. ruiter); https://orcid.org/0000-0001-7324-1735 (t. hopp). Clinical Biomechanics 110, 106117. doi:https://doi.org/10.1016/j.clinbiomech.2023.106117.

Salehi, Y., Giannacopoulos, D., 2022. Physgnn: A physics–driven graph neural network based model for predicting soft tissue deformation in image–guided neurosurgery. Advances in Neural Information Processing Systems 35, 37282–37296.

Samani, A., Bishop, J., Yaffe, M., Plewes, D., 2001. Biomechanical 3-d finite element modeling of the human breast using mri data. IEEE Transactions on Medical Imaging 20, 271–279. doi:10.1109/42.921476.

Sarno, A., Mettivier, G., di Franco, F., Varallo, A., Bliznakova, K., Hernandez, A.M., Boone, J.M., Russo, P., 2021a. Dataset of patient-derived 3d digital breast phantoms for research in breast computed tomography. URL: https://doi.org/10.5281/

zenodo.4529852, doi:10.5281/zenodo.4529852. data set.
Sarno, A., Mettivier, G., di Franco, F., Varallo, A., Bliznakova, K., Hernandez, A.M., Boone, J.M., Russo, P., 2021b. Dataset of patient-derived 3d digital breast phantoms for research in digital breast tomosynthesis and digital mammography. URL: https://doi.org/10.5281/zenodo.4515360, doi:10.5281/zenodo.4515360. data set.
Sarno, A., Mettivier, G., di Franco, F., Varallo, A., Bliznakova, K., Hernandez, A.M., Boone, J.M., Russo, P., 2021c. Dataset of patient-derived digital breast phantoms for in silico studies in breast computed tomography, digital breast tomosynthesis, and digital mammography. Medical Physics 48, 2682–2693. doi:https://doi.org/10.1002/mp.14826.
Schlömer, N., . pygalmesh: Python interface for CGAL's meshing tools. URL: https://github.com/nschloe/pygalmesh, doi:10.5281/zenodo.5564818.
Segars, W., Veress, A., Wells, J., Sturgeon, G., Kiarashi, N., Lo, J., Samei, E., Dobbins, J., 2014. Population of 100 realistic, patient-based computerized breast phantoms for multi-modality imaging research (br whiting & c. hoeschen, eds.; p. 90331x).
Sturgeon, G.M., Kiarashi, N., Lo, J.Y., Samei, E., Segars, W.P., 2016. Finite-element modeling of compression and gravity on a population of breast phantoms for multimodality imaging simulation. Medical Physics 43, 2207–2217. doi:https://doi.org/10.1118/1.4945275.
Tanner, C., Schnabel, J.A., Hill, D.L.G., Hawkes, D.J., Leach, M.O., Hose, D.R., 2006. Factors influencing the accuracy of biomechanical breast models. Medical Physics 33, 1758–1769. doi:https://doi.org/10.1118/1.2198315.
The CGAL Project, 2024. CGAL User and Reference Manual. 5.6.1 ed., CGAL Editorial Board. URL: https://doc.cgal.org/5.6.1/Manual/packages.html.
Wellman, P., Howe, R.D., Dalton, E., Kern, K.A., 1999. Breast tissue stiffness in compression is correlated to histological diagnosis. Harvard BioRobotics Laboratory Technical Report 1.
Zienkiewicz, O.C., Taylor, R.L., Zhu, J.Z., 2005. The finite element method: its basis and fundamentals. Elsevier.
Zyganitidis, C., Bliznakova, K., Pallikarakis, N., 2007. A novel simulation algorithm for soft tissue compression. Medical Biological Engineering Computing 45, 661–669. URL: https://doi.org/10.1007/s11517-007-0205-y, doi:10.1007/s11517-007-0205-y.

## Appendix A. Data generation

The data were prepared for training the PhysGNN model through the following steps:

- **Feature Extraction**: Features were derived from the mesh structure, including:
  - Elements IDs and their vertices indices
  - Initial nodal $x$, $y$, $z$ coordinates
  - Boundary condition IDs
  - Number of mesh nodes
  - Surface node indices
  - Normal directions ($x$, $y$, $z$) of the surface nodes

  Additionally, three other directions were randomly sampled from a hemisphere with a radius of 1.

- **Feature Matrix Construction**:
  - **Adjacency Matrix Creation**: An adjacency matrix was built to represent the connectivity of nodes in the FE mesh. Each element, defined by a set of nodes, connected all pairs of nodes within the same element. This matrix was processed to handle multiple directions and time steps by considering each combination of direction and time steps. For each element, if a pair of nodes were connected, their connection was adjusted for the current direction and time step by adding an offset to the node indices.
  - **Material and Boundary Conditions Extraction**: The material ID for nodes of each element was derived from the element ID, and the fixed nodal ID was derived from the boundary condition.

- **Time Steps and Directional Batches**: Features were constructed for 30-time steps, each corresponding to a deformation. Ten batches of directions were constructed:
  - Each surface node had four directions per batch.
  - Initial nodal coordinates, material IDs, and fixed nodal IDs were concatenated into a data frame, and repeated for the product of the number of directions and time steps.
  - An incremental force magnitude was applied at each time step for each surface node. This magnitude was calculated by dividing the total force magnitude by the product of the number of time steps and surface nodes. For each time step, the incremental force magnitude was multiplied by (*time step* + 1) and appended to a list of force magnitudes.
  - Directions ($x$, $y$, $z$) for each surface node were appended to a list, repeated for all nodal directions, ensuring correct distribution across nodes, time steps, and repetitions.

- **Force Matrix Construction**: Matrix $M$ was constructed with rows representing the force-length (number of time steps $\times$ number of directions $\times$ number of nodes) and four columns, filling in the force and its directions ($x$, $y$, $z$) for each surface node index. This data was then concatenated into the initial data frame.

- **Graph Initialization**: This construction was repeated for the ten batches of directions, resulting in:
  - Four graphs were initialized per time step, leading to 120 graphs generated from 30 time steps.

  - A graph indicator to determine node belonging across 120 graphs.
  - A graph label for labeling the 120 graphs.
  - A total of 1200 graphs (10 batches $\times$ 120 graphs per batch).

- **Feature Processing**: Based on the material ID and boundary condition, physical properties were assigned:

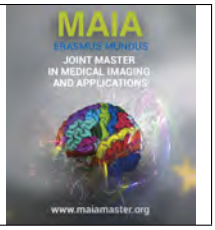
  - Nodes with free boundary conditions in fat tissue were assigned a value of 1, glandular tissue was assigned a value of 0.6, in skin tissue was assigned a value of 0.1.
  - Nodes with fixed boundary conditions were assigned a value of 0.

  The force magnitude was multiplied by the direction vectors ($x$, $y$, $z$) to obtain the force directions. These physical properties and force directions from each batch were saved (pickled).

- **Additional Feature Calculation**: Upon loading the dataset for training, additional features were calculated from the force directions in Cartesian coordinates, converting them into spherical coordinates.

Medical Imaging and Applications

Master Thesis, June 2024




# QEI-Net: A Deep learning-based automatic quality evaluation index for ASL CBF Maps

Xavier Beltran Urbano*,[1], Sudipto Dolui[1], John A Detre[1]

[1]*Detre Lab, Departments of Neurology and Radiology, University of Pennsylvania, USA*

**Abstract**

Arterial Spin Labeling (ASL) is a non-invasive magnetic resonance imaging (MRI) technique widely used for measuring cerebral blood flow (CBF). Compared to more conventional approaches, ASL offers several advantages, such as the absence of exogenous tracers and ionizing radiation, lower cost, flexibility of being acquired in routine MRI settings, and is the method of choice to measure CBF in large-scale multisite studies, particularly with repeated acquisitions. However, ASL data can be noisy, and hence quality control (QC) of ASL CBF maps is of particular importance for this modality. Manual QC is time-consuming, laborious, and subjective, highlighting the need for automated solutions. In this study, we proposed three novel deep learning (DL) models designed to provide automatic quality evaluation indices (QEIs) for ASL-derived CBF maps: 7FCN-QEI-Net, Reg-QEI-Net and MSC-QEI-Net. The resulting QEIs are designed to be continuous numbers in the range of 0 and 1. We also trained a deep learning algorithm (BC-Net) to provide a binarized decision about the quality of the CBF map, which indicates if the map should be kept or discarded from group analysis. Additionally, we also considered ensembles of the different networks. These approaches leverage advanced DL techniques to enhance feature representation and achieve superior performance compared to previous state-of-the-art methods.The models were trained on a diverse dataset that included 250 samples from multiple multisite studies. These samples were acquired using different protocols and were rated for quality by three raters, ensuring robustness and generalizability. Additionally, in a separate test set comprising 50 samples, all the deep learning strategies performed better than the current state-of-the-art method. The correlations between the automated QEIs and the average manual ratings were higher than the inter-rater correlations. We also derived and reported QEI thresholds for each method to binarize CBF maps into acceptable and unacceptable categories for each of the non-binarized methods. While the ensemble approaches perform slightly better, the Reg-QEI-Net provided comparable performance and is currently our recommended strategy. The results highlight the potential of DL models in automating and improving the QC process for ASL CBF maps, reducing reliance on manual assessments, minimizing subjectivity, and enhancing reproducibility and consistency across studies.

The code developed for this work is publicly available at: https://github.com/xavibeltranurbano/QEI-Net

*Keywords:* Arterial Spin Labeled, Deep Learning, CNN, Quality Assessment, FCN, Regression, Reg-QEI-Net, CBF

## 1. Introduction

The brain is one of the most highly perfused organs in the body, utilizing approximately 15% of the cardiac output and 20% of the total body oxygen (Jain et al. (2010)). Cerebral blood flow (CBF) is classically defined as the volume of blood flowing through a specific region of the brain tissue per unit time and is expressed in units of milliliters of blood per 100 gram of brain tissue per unit time (unit: ml/100g/min). It is an important physiological quantity of cerebrovascular health and provides an important biomarker for the latter. Changes in CBF correlate with various indicators of cerebrovascular disease, including white matter hyperintensities (Bernbaum et al. (2015)) and cerebral microbleeds (Gregg et al. (2015)). Additionally, it also

*Corresponding author
*Email address*: xavibeltranurbano00@gmail.com

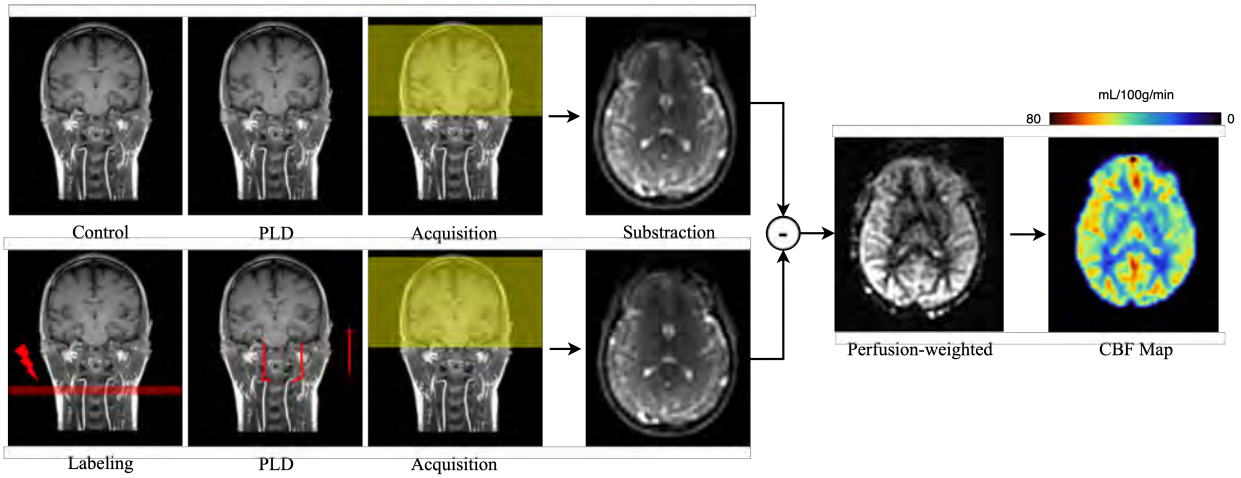



Figure 1: Sequential workflow for ASL CBF map acquisition. This diagram delineates the procedural stages, beginning with the acquisition of control images, followed by the application of labeling and post-labeling delay (PLD). Subsequent subtraction generates the perfusion-weighted images, which are then utilized to create the detailed CBF maps.

serves as a biomarker of functional neurodegeneration due to the strong association of changes in CBF with neural activity (Dolui et al. (2017a)), and therefore can potentially replace glucose metabolism measurements obtained using $^{18}$F-Fluorodeoxyglucose Positron Emission Tomography ($^{18}$F-FDG-PET)(Dolui et al. (2020)). CBF changes have been associated with the incidence and severity of dementia (Dolui et al. (2020);Dolui et al. (2017a);Binnewijzend et al. (2013);Wolk and Detre (2012)) and has been shown to be one of the earliest biomarkers to change in the Alzheimer's Disease continuum (Iturria-Medina et al. (2016);Dolui et al. (2024);Fazlollahi et al. (2020)). Moreover, CBF is potentially modifiable therapeutically and hence can be used to monitor treatment response (De La Torre (2013);Dolui et al. (2022)). Consequently, CBF measurement is considered very important in studies on healthy aging, cerebrovascular and neurodegenerative disease (Wolk and Detre (2012)).

### *1.1. Classical methods of measuring CBF*

Classical CBF is measured using a "diffusible" tracer that exchanges from the blood compartment to the tissue compartment, allowing CBF in ml/100g/min to be measured directly. The first CBF measurements in humans were made by Kety and Schmidt (Kety and Schmidt (1945)) by monitoring arteriovenous differences in nitrous oxide. The current "gold-standard" for CBF imaging in humans is $^{15}$O-PET scanning (Zhang et al. (2014);Herscovitch et al. (1983)), which utilizes radioactively labeled water as a perfusion tracer. Other diffusible tracer approaches used to measure CBF in humans include radioactive $^{133}$xenon (Lassen et al. (1981)) and stable xenon computed tomography (CT) (Yonas et al. (1991)). Related methods include accumulative radioactive tracers with single-photon emission computed tomography (SPECT) scanning, though agreement of these methods with $^{15}$O-PET is suboptimal (Ito et al. (2006)), and methods that use intravascular tracers such as perfusion CT (Koenig et al. (1998)) and dynamic susceptibility contrast (DSC) MRI (Rempp et al. (1994)). Intravascular tracer methods do not measure CBF directly but allow CBF to be inferred. All these methods require the administration of an exogenous tracer and exposure to ionizing radiation. Hence, they are at least somewhat invasive and can be difficult to administer to clinically vulnerable population groups, including the elderly, infants, and individuals with renal impairments. Moreover, using such methods to track CBF changes in healthy aging and in drug studies can be problematic, as these studies require serial measurements with repeated exposure to tracers or ionizing radiation and associated costs.

### *1.2. Arterial Spin Labeled (ASL) perfusion MRI*

ASL is a non-invasive magnetic resonance imaging (MRI) technique for measuring tissue perfusion by magnetically labeling arterial blood water as an endogenous tracer (Detre et al. (1992)). Since its inception in 1992 (Detre et al. (1992);Williams et al. (1992)), ASL has been increasingly included in multisite research studies of brain health. Compared to other techniques for measuring cerebral perfusion, ASL offers advantages due to its non-invasive nature and the absence of exogenous radioactive and potentially harmful contrast agents. Furthermore, because MRI does not involve ionizing radiation, this method can be used repeatedly, for example, to assess the effects of drugs or to assess longitudinal changes in cerebral perfusion. Finally, ASL can be acquired as a part of routine MRI,

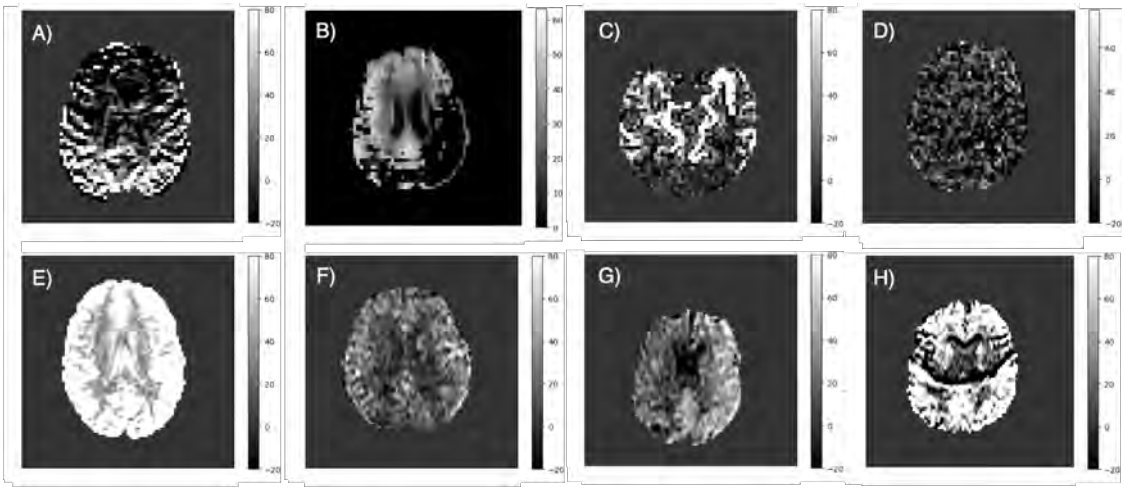



Figure 2: Examples of different sources of artifacts in ASL CBF maps. A) Motion Artifact B) Clipping Artifact C) Transit Artifact D) Low SNR E) High CBF Values F) Low CBF Values G) Probable Label Asymmetry H) Fat Shift Artifact.

which is almost universally acquired in research studies of brain disorders. ASL has been validated against other established modalities for measuring CBF (Ewing et al. (2005);Heijtel et al. (2014);Ye et al. (2000a)). Its use also extends beyond the brain studies and is being applied to other organs, including the kidneys, lungs, heart, placenta, eye, liver, pancreas, and muscle (Taso et al. (2023)). ASL MRI has also been translated to clinical use.

### *1.3. ASL MRI Data Acquisition*

The acquisition of ASL MRI data involves magnetically labeling inflowing protons of proximal arterial blood water. For brain perfusion, labeling typically occurs in the neck, where blood flows through the internal carotid and the vertebral arteries that supply blood to the brain (see Figure 1). After waiting for a brief period (post-labeling delay) to allow the flow of the labeled blood to reach brain microvasculature and tissue, a brain MRI (labeled image) is acquired. A "control" brain image is also obtained with a sham labeling procedure that does not magnetically label blood water. The difference between the control and label image is proportional to CBF and is converted to absolute CBF quantification using a proton density image with appropriate models and assumptions (Alsop et al. (2015);Buxton et al. (1998)). The control-label difference is a small percentage of the background signal, which results in a low signal-to-noise-ratio (SNR) in the CBF images. Additionally, subject motion, suboptimal choice of imaging parameters, and other non-idealities inherent to MRI scanners can lead to severe artifacts (Dolui et al. (2017b);Li et al. (2018)) (see Figure 2). This can be partially mitigated by averaging multiple control-label pairs, using advanced signal processing strategies, and using background suppression (BS) of static brain water. BS increases the difference image by 3-10 times (Dolui et al. (2019);Maleki et al. (2012);Ye et al. (2000b)). Nevertheless, a noticeable amount of artifact might remain in the resulting CBF image.

### *1.4. ASL Labeling Methods*

Ever since its establishment in 1992, several ASL protocols have been devised and used, which primarily differ in labeling and signal readout strategies (see Figure 3). The classical method invented in 1992, which was referred to as Continuous ASL (CASL) (Detre et al. (1992)), continuously saturates or inverts arterial blood water at the neck for several seconds. However, modern human MRI scanners utilizing whole-body radiofrequency (RF) amplifiers are not capable of continuous RF excitation. Pulsed ASL (PASL) instantly labels a thick slab in the neck, and is compatible with body RF excitation, though the method suffers from lower SNR compared to CASL. The current recommended labeling strategy is pseudo-continuous labeling (PCASL), which employs a series of short RF pulses to mimic continuous labeling. ASL type can also vary based on the duration of the post labeling delay (PLD) – a longer post labeling delay can ensure delivery of the labeled blood to the brain tissue, though at the expense of reduced SNR since the magnetic label decays rapidly. A series of ASL images acquired with different labeling and/or PLDs can also be combined to obtain a CBF map, allowing more accurate modeling of regional CBF values (Woods et al. (2024)). Finally, ASL image quality can vary based on the type of image readout. Echo-planar imaging (EPI) was initially the preferred choice because of speed and sensitivity, though it is being slowly replaced by 3D imaging (GRASE or SPIRAL) optimally combined with BS of static brain tissue. Notably, several other variants of ASL exist; for example, velocity selective ASL (VASL) is an emerging method that labels the arterial blood water close to the imaging site

instead of the neck (Qin et al. (2022)).

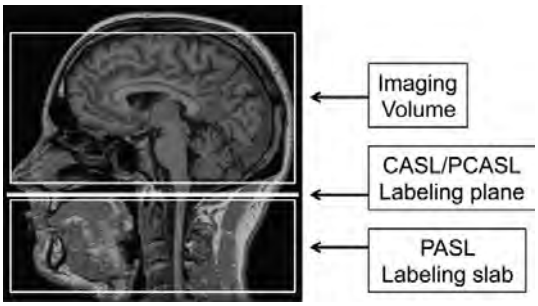



Figure 3: Schematic diagram of imaging and labeling regions for CASL/PCASL and PASL. In CASL/PCASL, labeling occurs as blood flows through a single labeling plane, while in PASL, a slab of tissue, including arterial blood, is labeled (Alsop et al., 2015).

### *1.5. Artifacts in ASL MRI and the need for an automated quality evaluation index (QEI)*

In recent years, ASL has gained popularity among perfusion imaging modalities for its use in research settings, largely due to its potential as a biomarker of cerebrovascular health and brain function and its ability to be acquired in routine MRI settings. Despite recent advancements in improving the quality of ASL images, the resulting CBF maps can still be contaminated by artifacts. The most significant source of artifact is physiological noise due to motion, particularly in non-compliant subjects or in patients who have difficulty staying still during the scan. Because the control/label difference represents only a small percentage of the background signal, any variability in the background signal due to motion can dominate the difference signal, leading to large errors that are often not removed during averaging. Retrospective motion correction techniques are generally used to account for bulk motion, but such techniques cannot correct for variation in intensities occurring during the image readout (Friston et al. (1996);Power et al. (2012)). Motion effects are less visible, though still present, in acquisitions using BS of static signal (Ye et al. (2000a);Fernandez-Seara et al. (2005);Maleki et al. (2012)). Artifacts can also result from an incorrect or suboptimal choice of acquisition parameters. For example, an insufficiently long PLD results in labeled blood remaining in large arteries rather than in the microvasculature or parenchyma, an effect known as transit time artifact that affects both BS and non-BS acquisitions. Other problems inherent to MR imaging, such as thermal noise, chemical shift artifacts, and clipping of signals, can produce errors and artifacts in the resulting CBF maps. For clinical research, another concern is that the number of corrupted ASL CBF images may increase with disease severity, as previously found in the AD continuum (Moonen et al. (2020)), making QC a more prominent need in these clinical applications.

Because of potential artifacts in the ASL derived CBF maps, QC is critical for clinical research of ASL MRI to exclude CBF maps of poor quality that can reduce sensitivity to biological effects of interest. Current QC heavily depends on manual assessment, which is time consuming, laborious, and subjective, and therefore not reproducible and generalizable, especially for large-scale multisite studies. Therefore, there is a critical need for a robust and reliable automated quality evaluation index (QEI) that can objectively assess the quality of ASL CBF scans. This QEI could also potentially facilitate real-time feedback during scanning, allowing for immediate adjustments and thereby improving the overall quality of the acquired images.

### *1.6. Deep Learning*

Deep Learning (DL), a subtype of machine learning, provides astonishing performance compared to other state-of-the-art computational methods across various approaches (Bengio et al. (2013);Deng and Dong (2014);Lecun et al. (2015);Litjens et al. (2017)), including medical imaging. Initially introduced for image classification in computer vision (Krizhevsky et al. (2012)), DL is now extensively employed to tackle complex problems that analytical methods or traditional machine learning cannot solve. DL networks are motivated by the neuronal visual processing pathway, where a visual observation is hierarchically processed along multiple layers of neurons and eventually abstracted to different top-level features. Multi-layer artificial neural networks were proposed decades ago to mimic this complex learning process, but their use only became practical with the advent of powerful graphical processing units (GPUs) capable of massively parallel computing (Bengio et al. (2013);Deng and Dong (2014);Lecun et al. (2015);Litjens et al. (2017)). Deep networks are commonly trained with references; this supervised learning is equivalent to nonlinear data fitting. While traditional data fitting is based on a weighted sum of well characterized base functions, DL is based on the weighted sum of the output of a hierarchical network consisting of multiple layers of computing units (artificial neurons).

### *1.7. Contribution of this work*

In this work, we aimed to tackle the challenge of providing an automatic and robust QC method for ASL-derived CBF maps by leveraging DL. We explored multiple strategies to derive this metric, including both the use of predetermined features and the entire CBF map for automatic feature extraction. We then compared their performances, demonstrating their superiority over previous approaches.

The specific contributions of this work include the development of the following DL-based methods to obtain a QEI of raw CBF maps:

- A feature-based regression model, for which we extracted 7 predetermined features to train a fully connected network (named 7-FCN-QEI-Net).
- A 3D DL-based regression model (named Reg-QEI-Net).
- A 3D multi-stage classification model (named MSC-QEI-Net).
- A 3D binary classification model (named BC-Net).
- Three ensemble methods of the best performing algorithms.

An extensive comparison of these new approaches with the current state-of-the-art method was performed, providing insights into their relative performances and improvements.

## 2. State of the art

### *2.1. DL-based regression approaches for neuroimaging*

Since deep learning models first made their mark on neuroimaging in 2014 (Plis et al. (2014)), there has been an exponential increase in research within the field. This remarkable growth can be attributed to two main factors: the increasing availability of data and the improvement of computational resources such as GPUs. Thanks to these advancements, deep learning has emerged as a leading approach in medical imaging research, with segmentation and classification tasks ranking at the forefront of the most explored areas. However, regression tasks, which aim to predict a continuous outcome, have received comparatively less attention due to their perceived complexity. Consequently, several studies, such as that by (Peng et al. (2021)), have opted to recast the initial regression challenge into a classification problem by discretizing the continuum of values into distinct bins, treated as independent classes during training. (Leonardsen et al. (2022)) delve into a comparative analysis of both methodologies, focusing on predicting brain age from structural MRI scans. Employing a 3D Convolutional Neural Network (CNN) architecture with six convolutional blocks, the study experimented with both approaches by merely altering the last dense layer and meticulously fine-tuning the hyperparameters for each approach. Although the outcomes on the test set were comparably effective for both approaches, the regression method demonstrated markedly superior generalization capabilities on an unseen dataset, thereby underscoring its enhanced potential for broader applicability. In line with these findings, recent studies highlight the increasing sophistication of deep regression models tailored for neuroimaging data. For instance, (He et al. (2022)) introduced deep relation learning, which utilizes a novel approach by considering multiple relational aspects between neuroimaging inputs to enhance regression performance in age estimation tasks. By leveraging deep neural networks to capture complex and non-linear interactions, this method provides a more nuanced understanding and robust predictions than traditional methods.

### *2.2. Deep Learning-based approaches for ASL MRI*

In recent years, there have also been notable advancements in the utilization of DL for ASL MRI, resulting in considerable improvements when dealing with certain intrinsic difficulties associated with this image modality, including its lengthy acquisition periods, inadequate SNR, and low spatial and temporal resolution. In their study, (Kim et al. (2018)) reported significant advancements in the quality of ASL MRI images using CNNs that surpassed those created by traditional averaging techniques. Building on these improvements in imaging techniques, the application of transfer learning has demonstrated potential for augmenting sensitivity, especially in clinical contexts involving AD. For instance, (Zhang et al. (2022)) highlighted the efficacy of applying transfer learning from healthy subjects to ASL perfusion MRI models. This approach significantly increased the sensitivity of detection methods for AD, illustrating how advances in deep learning could be specifically tailored to improve diagnostic processes. The investigation conducted by (Xie et al. (2020)) presented an innovative DL-based ASL MRI denoising algorithm that improved the SNR of CBF images and enabled a 75% reduction in acquisition time while maintaining the integrity of the measurements. Similarly, (Gong et al. (2020)) introduced a DL algorithm for denoising ASL MRI that combines CNNs and mutual information from multiple tissue contrasts in ASL acquisition. This approach demonstrated superior performance over traditional and standard deep learning-based denoising methods by significantly enhancing image quality.

### *2.3. Quality index of ASL CBF maps*

As previously stated, QC of ASL CBF maps through visual inspection is a labor-intensive process that requires significant expertise. This method is also prone to user bias and subjectivity, particularly when applied to large sample sizes. The work in (Fallatah et al. (2018)) introduced a well-characterized dual-component scoring system that evaluates the image quality based on visual contrast and artifact detection and thus reduces the subjectivity of the rating system. This system, validated across multiple raters, has demonstrated high reproducibility and the ability to effectively discriminate between high- and low-quality clinical scans, offering a reliable threshold for clinical acceptability; however, it still suffers from most of the drawbacks of manual rating.

Parallel to these manual evaluation strategies, there have been efforts to automate quality assessments. For

instance, (Li et al. (2019)) developed ASLMRICloud, an online platform that facilitates the processing of ASL MRI data. Among other features, ASLMRICloud enables the calculation of a quality index by analyzing and averaging the voxelwise temporal standard error (SNR) across the CBF time series obtained from the repeated acquisitions of the multiple control/label pairs. However, this approach cannot assess systematic artifacts that are consistent in the time series, such as those caused by short PLD. Moreover, it cannot be applied to datasets that include only one output volume of the average control-labeled difference image rather than the control-labeled image time series (e.g., product ASL on a GE MRI scanner). Finally, temporal standard error considers the quality of the raw data instead of the final CBF map, which can be of improved quality through the application of signal processing strategies.

The most recently published contribution to the development of an automated QEI for ASL CBF maps was made by (Dolui et al. (2024)). This novel QEI assigns a continuous value between 0 and 1 to each CBF map, with higher values indicating a superior quality of the CBF map. The algorithm used predefined features to train a model against human rating, where the features were chosen to replicate the meticulous visual inspections usually performed by experts during manual QC. The computational features integrated into the QEI methodology involve:

- **Structural Similarity:** The QEI considers the similarity between the brain structure and CBF maps, acknowledging the natural correlation between structure and function. This feature is calculated by constructing a structural pseudo-CBF (spCBF) map, utilizing a weighted sum of tissue probability maps to reflect the higher CBF in gray matter (GM) compared to white matter (WM). The Pearson correlation between the spCBF map and the original CBF map was used as a feature in the QEI derivation.
- **Spatial Variability:** Although CBF differs among tissue types, unusual spatial variability might suggest the presence of artifacts, such as those from motion or inadequate PLD (see examples in Figure 4). Therefore, to accurately reflect these variations, QEI integrates a dispersion index (DI) for CBF values across GM, WM, and cerebrospinal fluid (CSF) masks, normalized by the mean GM CBF.
- **Negative GM CBF:** Given that physiological CBF should be positive, the QEI incorporates the proportion of GM voxels showing negative CBF values, since those voxels represent non-physiological artifact-affected measures.

The final QEI was performed by fitting these features separately to human ratings of 101 CBF maps, and subsequently performing a geometric average of the fits corresponding to each feature as follows:

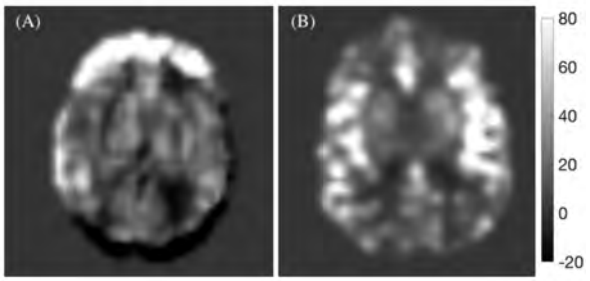



Figure 4: Examples of large spatial variability in ASL derived CBF (A) due to motion or (B) the post-labeling delay (150ms) being significantly shorter than the arterial transit time resulting in labeled signal retained in the arteries instead of the tissue parenchyma while imaging (Dolui et al. (2024)).

$$QEI = \sqrt[3]{\left(1 - e^{-3p_{ss}^{2.4}}\right) e^{-\left(0.1DI^{0.9} + 2.8p_{nGMCBF}^{0.5}\right)}} \quad (1)$$

where

- $p_{ss}$ is the structural similarity.
- $DI$ is the spatial variability.
- $p_{nGMCBF}$ is the proportion of negative voxels in GM CBF maps.

This method showed similar agreement to inter-rater reliability, improved statistical analyses, and performed better than the method developed by (Li et al. (2019)). Consequently, it is recognized as the state-of-the-art in automatic QEI for ASL CBF Maps. Therefore, we have used this study as a benchmark to compare the various approaches presented in this work.

## 3. Material and methods

### *3.1. Datasets*

In this study, a dataset comprising 250 samples was utilized to train the different models. The samples were collected from several large, multisite studies that utilized diverse ASL acquisition protocols, as detailed in Table 1. The ratings of the ASL CBF data were meticulously assessed by three expert raters: John A. Detre, Sudipto Dolui, and Ze Wang. Dr. Detre, the inventor of ASL, has over 30 years of experience, while Dr. Dolui and Dr. Wang each have more than 10 years of experience with this technique. Their extensive experience in ASL CBF quality assurance ensures the dataset's reliability and validity. Additionally, a separate set of 50 CBF maps rated by Dr. Detre and Dr. Dolui was used as the test set to assess the performance of the algorithms on unseen data. All the data used in this project have been acquired using Siemens MRI scanners.

Table 1: Information of the different datasets used in this work.

| Dataset | Protocol | Sample Size |
|---|---|---|
| Alzheimer's Disease Neuroimaging Initiative (ADNI) (Wang et al. (2013)) | 2D PASL | 79 |
| Multi-Ethnic Study of Atherosclerosis (MESA) (Austin et al. (2024)) | 3D BS PCASL | 57 |
| Systolic Blood Pressure Intervention Trial (SPRINT) (Dolui et al. (2022)) | 2D PCASL | 49 |
| Coronary Artery Risk Development in Young Adults (CARDIA) (Dolui et al. (2016)) | 2D PCASL | 25 |
| National Alzheimer's Coordinating Center (NACC) (Dolui et al. (2019)) | 3D BS PCASL | 34 |
| Vascular Contributions to Cognitive Impairment and Dementia (VCID) (Sadaghiani et al. (2023)) | 3D BS PCASL | 6 |

To ensure consistency in the evaluation process across different raters, specific guidelines were established and followed (see Figure 5). These guidelines are defined below:

- **Unacceptable (rating 1):** CBF map is severely degraded by artifacts and is uninterpretable.
- **Poor (rating 2):** CBF map has one or more major artifacts, but can still potentially yield useful information.
- **Average (rating 3):** Acceptable quality CBF map with minor artifacts that do not significantly reduce information value.
- **Excellent (rating 4):** High quality CBF map without artifacts.

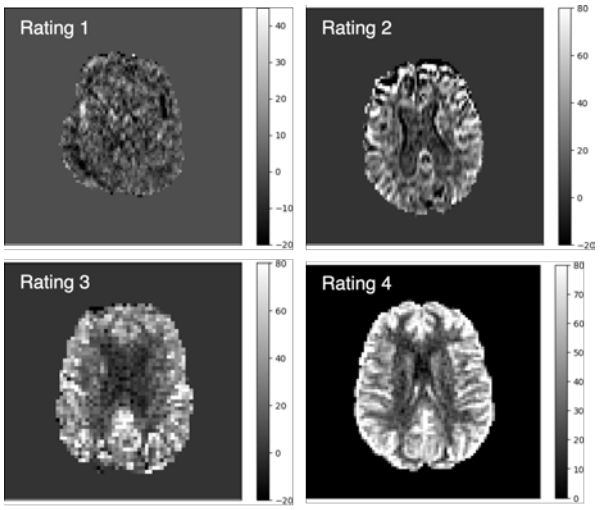



Figure 5: Examples of a distinct case for each rating value.

In the regression-based approaches (mentioned in the introduction and in more detail below), we averaged the ratings to obtain a composite rating score and also to increase the reliability of the measures. Furthermore, we wanted the final QEI to be in the [0,1] range and hence normalized the ratings between 0 and 1. To facilitate the rating process, a specialized tool was developed, as outlined in **Appendix A**.

### *3.2. Dataset Partitioning*

To validate the proposed approaches, we employed a 5-fold cross-validation (CV) strategy. Thus, in each fold, 80 percent of the data was used to train the model, and the remaining 20 percent was kept as a validation set. Finally, as previously mentioned, we tested our models using a test set consisting of 50 samples.

### *3.3. Preprocessing*

The CBF maps were derived from ASL data using standard processing strategies (Alsop et al. (2015)). For the purpose of developing the QEI, additional preprocessing was required (see Figure 6). We have followed two different DL strategies, a FCN based on predetermined features and CNNs using the CBF images. For the former approach, two preprocessing steps were applied:

- Generation of binary masks corresponding to GM, WM and CSF to extract CBF signal in the regions.
- Smoothing of the CBF images using a 5 mm isotropic kernel. A similar approach was used by (Dolui et al. (2024)) to extract features from the CBF maps.

For the CNN approaches (Reg-QEI-Net, MSC-QEI-Net, and BC-Net), we used the SimpleITK library to perform an affine transformation, resampling the dimensions and spacing of the images to a uniform size of 64x64x32. This step accounted for variations in image sizes acquired across different studies and protocols. After resampling, the images were intensity-clipped to the range [-10, 80] and subsequently normalized to a range of [0, 1] before being fed into the network.

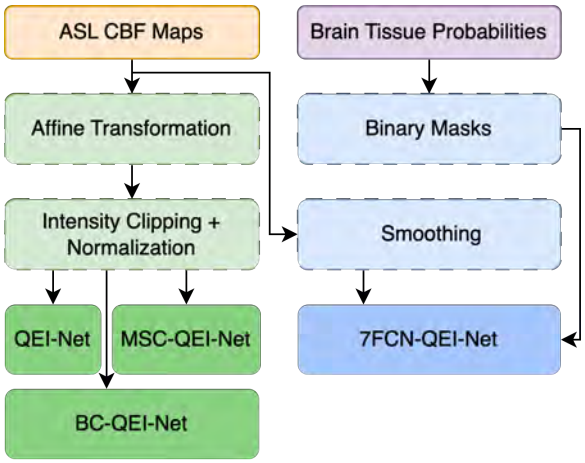



Figure 6: Workflow of the preprocessing pipeline.

### 3.4. Data Augmentation

Data augmentation techniques are methods used to artificially increase the variability of a dataset by applying various transformations to the original data. These transformations enhance the generalization capabilities of CNN models by exposing them to a wider range of variations. In this work, we used random vertical and horizontal flips, as well as rotations between -5 to 5 degrees.

### 3.5. Deep Learning models

#### 3.5.1. 7- Feature-based FCN model (7-FCN-QEI-Net)

As previously stated, (Dolui et al. (2024)) introduced a novel algorithm that utilizes three key features commonly employed in manual QC of ASL CBF maps to provide a QEI. While this method achieved high performance and set a new benchmark in the field, its capability is likely constrained by the limited number of features. In our research, we build upon that foundational work by proposing the integration of four additional features. These features are as follows:

- **SNR:** For this feature, we have computed the spatial SNR as the ratio of the GM CBF to the standard deviation of the signal in CSF CBF.
- **Summary Statistics:** Several statistics are calculated from the GM and WM of CBF Maps. They consist of the mean, the inverse of the standard deviation, and 5th and 95th percentiles of kurtosis.
- **Shannon Entropy:** To measure the ghosting and blurring induced by head motion, we have computed the Shannon entropy. The inverse of this measure is used as a feature for our model.
- **Spatial Gradients:** In ASL CBF maps, there can be differences in intensities along the three axes due to possible intensity variation or incorrect application of model equations. The variance of the inverse of CBF map gradients along each spatial dimension is then used as a feature for our model.

After computing these features, they are combined with the features from (Dolui et al. (2024)) and used as input for an FCN architecture (named **7-FCN-QEI-Net**) comprising of seven fully connected layers (FCL) with [64,256,512,256,64,16,1] neurons in each layer, respectively. In the last layer of this network, a sigmoid activation function is used to predict a continuous value constrained between [0,1]. Finally, squared error (SE, defined in section 3.7 below) was designated as the principal metric for this project, and thus, Mean Squared Error (MSE) was used as the loss function for the training of this model. An example of this network is presented in Figure 7.

#### 3.5.2. Deep learning-based regression model (Reg-QEI-Net)

Next, instead of the manual feature extraction used in the 7FCN-QEI-Net, we opted for data-driven approaches using CNNs where the CBF maps were used as input. These methods do not require a segmented image of different brain tissues, making them effective even when a structural image necessary for accurate segmentation is unavailable. This technique involves a sophisticated deep-learning based regression model, which we have named **Reg-QEI-Net**.

Drawing inspiration from the 3D VGG architecture delineated by (Simonyan and Zisserman (2014)), we have incorporated several tailored modifications. The presented network, illustrated in Figure 7, is structured into four convolutional blocks, each augmented with residual connections to mitigate the vanishing gradient problem (see Figure 8). After the first three blocks, max pooling layers with a pooling size of 2 are employed for downsampling each channel. The network concludes with a series of three FCL, culminating in a final neuron activated by a sigmoid function. For better weight initialization, we utilized Glorot's initialization method (Glorot and Bengio (2010)), which ensures the variance of activations remains consistent across every layer, preventing the gradient from exploding or vanishing. The Adam optimization algorithm was used with an initial learning rate of 0.0001. Moreover, a batch size of 32 samples and a learning rate decay strategy were applied, with a decay factor of 0.1 and a patience threshold of 15 epochs. Although the training was initially set to run for 400 epochs, an early stopping mechanism with a patience parameter of 60 epochs was implemented to prevent overfitting. Additionally, a dropout rate of 20% was applied after the fully connected layers to further prevent overfitting. Finally, similar to the 7FCN-QEI-Net approach, MSE was used as the loss function for training this model.

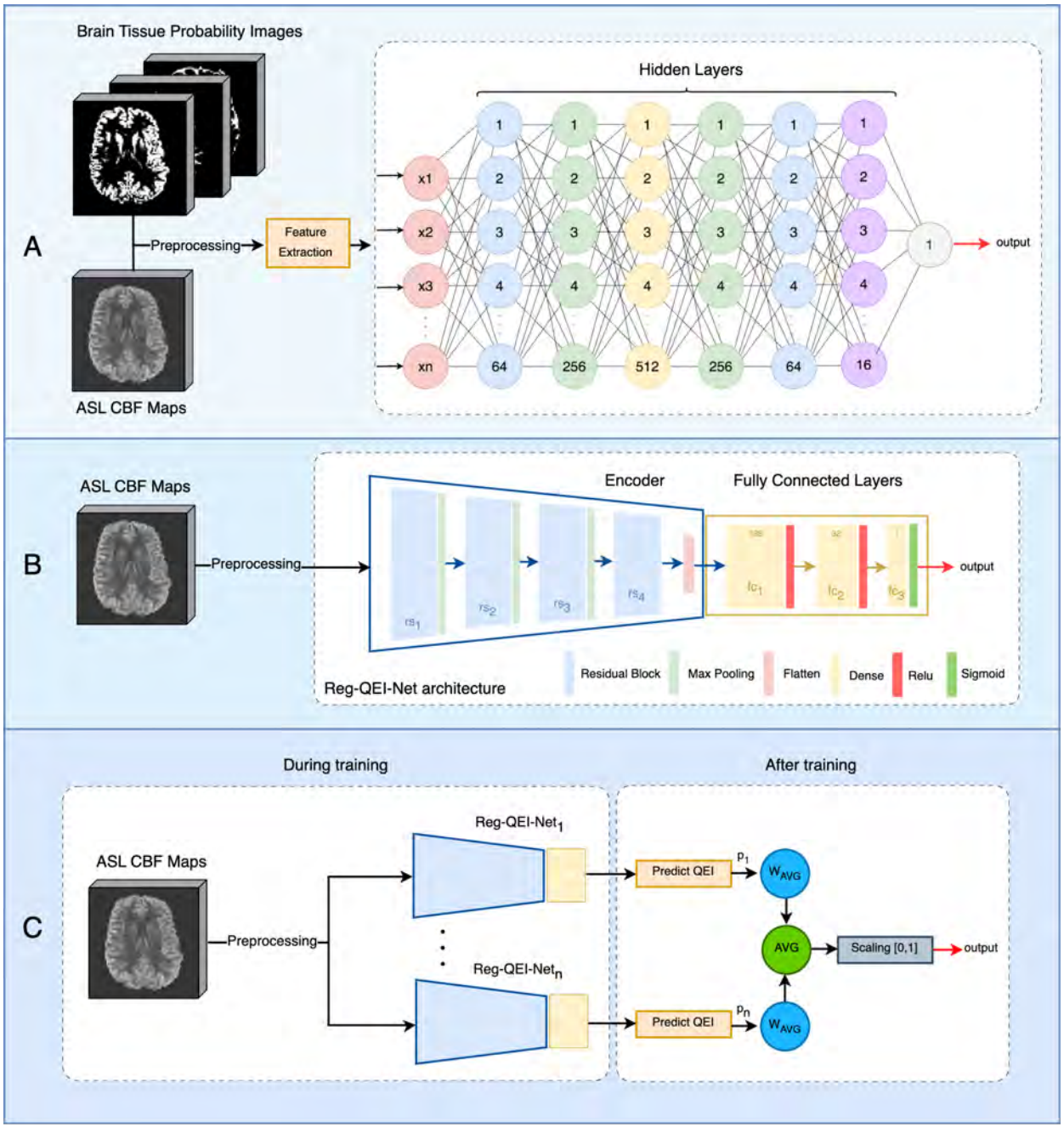



Figure 7: Schematic of the different deep learning pipelines implemented in this work. A) Feature-Based approach (FCN-QEI-Net) B) Regression approach (Reg-QEI-Net) C) Multi-Stage Classification approach (MSC-QEI-Net).

### *3.5.3. A 3D Multi-Stage Classification Model (MSC-QEI-Net)*

As delineated in Section 2.1, current advancements in deep learning-based regression models typically reformulate the regression problem as a classification task. This is achieved by discretizing the prediction range into distinct intervals, each representing a unique label. While this technique has been shown to enhance the efficacy of regression methods, it does have a substantial drawback: the precision is dependent on the number of intervals (bins) that are defined. An increased number of bins can yield higher precision, but it also intensifies the data imbalance among the bins. To address these challenges, we propose a multi-stage classification methodology named **MSC-QEI-Net**. This novel framework diverges from the aforementioned methods, which are focused on converting a regression task into a classification one by dividing the output into bins. Instead, MSC-

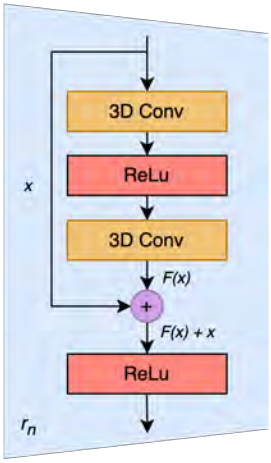



Figure 8: Schematic of the Residual Block used in this study. In the diagram, $r_n$ indicates the sequence number of the block, reflecting their multiple uses throughout the model.

QEI-Net comprises a series of multi-label classification networks, each corresponding to an individual rater's assessments within the dataset. The network used to perform this classification is based on the one presented in Section 3.5.2 with some minor changes. In this architecture, since we want to perform multi-label classification instead of regression, the last FCL contains 4 neurons, corresponding to each of the labels of the classification. In line with this modification, the softmax activation function, which is widely used for multi-label classification tasks, was utilized as the activation function of this layer. For both optimization and training, we applied similar hyperparameters to those previously used in the Reg-QEI-Net model. For the loss function, however, we opted for Focal Categorical Crossentropy loss, a prevalent choice in multiclass classification tasks with imbalanced data.

After training the network, we compute the weighted average of the prediction by following the formula delineated in Equation 2.

$$\text{Weighted Average Prediction} = \sum_{i=1}^{n} (p_i \cdot i) \qquad (2)$$

Where:

- $n$ is the number of classes.
- $p_i$ is the prediction score for the $i$-th class.
- $i$ is the class label, ranging from 1 to $n$.

Then, by aggregating the outputs of these networks and subsequently normalizing them, the system synthesizes a continuous value within the [0,1] range, representing the QEI of the image.

*3.5.4. A 3D Binary Classification Network (BC-Net)*

One of the main objectives of this project is to develop a robust method for discarding unacceptable CBF maps, which can be framed as a binary classification problem instead of assigning a continuous number defining the quality. Therefore, we also implemented a 3D binary classification approach named **BC-Net**. To do so, we have first binarized the expert ratings by following these criteria:

- **Unacceptable Quality (0):** if any of the raters gave a rating of 1 to the image.
- **Acceptable Quality (1):** otherwise.

Furthermore, we used the same parameters and architecture as the Reg-QEI-Net methodology described in Section 3.5.2. However, some minor adjustments were made to optimize the network. The main difference lies in the ground truth used to train the network. For Reg-QEI-Net, we used continuous values within the range [0,1], whereas for BC-QEI-Net, we used binary decision values explained above. For this reason, we utilized a binay cross-entropy loss function and a sigmoid activation function in its final FCL. The output of the BC-Net falls within the range of 0 to 1, representing the probability that a given sample is of acceptable quality.

*3.5.5. Additional Experiments*

Various combinations of the previous methods (Reg-QEI-Net, 7FCN-QEI-Net, and MSC-QEI-Net), that could potentially result in a better model, were also studied. BC-Net was not used in the combination since it represents a binary decision while other outputs a QEI value. The different combination methods are as follows:

- **Ensemble 1:** This is the simplest ensemble method, which consists of averaging the predictions from each of the networks.
- **Ensemble 2:** In this method, we calculate the weighted average of the predictions. To calculate the weights of each method, we have trained a function that optimizes the weights assigned to the different models to minimize the MSE between the ratings and the predictions.
- **Ensemble 3:** This method utilizes stacking, an ensemble technique that combines the predictions of multiple base models to enhance predictive performance. In this approach, the predictions from the QEI models serve as input features for a meta-model, which was trained using a 5-Fold CV with a linear regression algorithm that learns to make the final prediction by leveraging the strengths and mitigating the weaknesses of the individual models.

To limit the number of ensembles, only the best-performing models (Reg-QEI-Net and 7FCN-QEI-Net, see Section 4) were used. After training Ensemble 2 on the validation data, the resulting weights assigned to Reg-QEI-Net and 7FCN-QEI-Net were 0.663 and 0.337, respectively. These weights were then used to compute the weighted average of the predictions on the test data. Similarly, after training the linear regression models on the validation set, these models were subsequently applied to the test set.

### *3.6. Gradient-weighted Class Activation Mapping (Grad-CAM) and Heatmap Generation*

The QEI developed from the above approaches provides a summary metric for assessing the overall quality of the entire image. However, when the quality is not perfect, the QEI only indicates the presence of the artifacts in the image, without providing information about the location of the artifact. This is important information in region of interest (ROI) analysis as the mean CBF in the corresponding ROI can be contaminated by artifacts, although the overall CBF map might pass the QEI threshold, and that can subsequently bias the analysis. To visualize where the networks are focusing their attention, or in other words, which region of the image is contributing most to the QEI, we have implemented Gradient-weighted Class Activation Mapping (Grad-CAM) Selvaraju et al. (2017). Grad-CAM leverages the gradients flowing into a chosen convolutional layer to generate a localization map, or heatmap, which highlights the important regions in the input image. This technique provides a visual explanation for the model's predictions by identifying the areas in the brain images that contribute the most to the network's decision-making process. For our implementation, we have utilized the Reg-QEI-Net model to generate the heatmap. Among all the convolutional layers of the network, we utilized the 5th 3D convolutional layer, which is located in the third residual block. This decision was made because this intermediate layer provides a balance between low-level feature extraction and high-level semantic information, making it ideal for generating detailed and informative heatmaps.

### *3.7. Algorithm Evaluation Metrics*

To assess the performance of the algorithms, we computed the SE between the average manual ratings and the automated QEI for each CBF map, as defined below.

$$\mathrm{SE}_i = (\hat{r}_i - \overline{r}_{\mathrm{norm},i})^2 \tag{3}$$

with:

- $\overline{r}_{\mathrm{norm}}$: Normalized average rating of the experts.
- $\hat{r}_i$: Predicted rating.

In addition to that, we also reported the Pearson's correlation (PC) coefficient between the automated QEI and the average human rating and compared that to the correlation between the raters. Finally, dividing the data as unacceptable and acceptable as described in Section 3.5.4, we computed the receiver operating characteristic (ROC) curve and the area under the curve (AUC). To establish a QEI threshold, we have calculated the Youden Index (YI), as introduced by (Ruopp et al. (2008)). The YI is a statistical measure that aims to maximize both sensitivity and specificity. By computing the euclidean distance between all points of the ROC curve and the ideal point located at the coordinates [0,1], the YI identifies the best operating point in the curve. Thereafter, we computed sensitivity and specificity based on that threshold.

### *3.8. Computational resources*

The models were implemented using Python version 3.10.12 and TensorFlow version 2.16.1. The experiments were conducted on Google Cloud Platform (GCP) using a 64-bit GNU/Linux operating system (Ubuntu 22.04.04). The server was equipped with two Intel Xeon CPUs (2.30GHz), 8 GB of RAM, and a Tesla T4 GPU with 16 GB of memory, utilizing CUDA 12.4 for the experiments.

## 4. Results

### *4.1. Algorithm Evaluation Metrics*

Table 2 shows the mean, standard deviation, median, and IQR of the SE of the validation set (obtained from the 5-fold CV strategy), while Table 3 shows the same for the test set. Figure 10(a) and Figure 10(b) present the violin plots for the same. Table 2 and Table 3 also list the PC coefficients with the average expert ratings. Notably, the PC coefficient for the 250 samples used for training is 0.85 between Dolui and Detre, 0.84 between Dolui and Wang, and 0.80 between Detre and Wang. Furthermore, the correlation coefficient between Dolui and Detre was 0.77 for the test data set. In each case, the agreement between the raters was lower than the agreement between the average rating and the automated methods.

Additionally, Table 2 and Table 3 show the AUC, sensitivity, and specificity as detailed in Section 3.7. Note that the YI was based on the validation set and hence has not been presented in the table related to the test set. Figure 11(a) and Figure 11(b) show the ROC for the validation and the test sets. As expected, the performance of the test set was slightly worse than the validation set based on all the metrics. Although all the algorithms provided comparable performance, with the

Table 2: Comparison of the current state-of-the-art in the field of QEI of ASL CBF Maps (Dolui et al. (2024)) with the different QEI methods presented in this study using the validation data set.

| Method | MSE ± std SE | Median of SE (IQR) | PC Coefficient | AUC | Sensitivity | Specificity | YI |
|---|---|---|---|---|---|---|---|
| Dolui et al. 2024 QEI | 0.02160 ± 0.03184 | **0.00416 (0.01416)** | 0.943 | 0.948 | 0.904 | 0.922 | 0.457 |
| 7FCN-QEI-Net | 0.01646 ± 0.02986 | 0.01044 (0.02562) | 0.903 | 0.950 | **0.911** | 0.922 | 0.325 |
| Reg-QEI-Net | 0.01251 ± 0.02213 | 0.00611 (0.01556) | 0.923 | 0.958 | 0.815 | **0.965** | 0.461 |
| MSC-QEI-Net | 0.02123 ± 0.02579 | 0.01348 (0.02287) | 0.921 | 0.941 | 0.822 | 0.930 | 0.419 |
| BC-Net | - | - | - | 0.940 | 0.889 | 0.852 | 0.614 |
| Ensemble 1 | 0.01144 ± 0.02008 | 0.00505 (0.01124) | 0.947 | 0.963 | 0.889 | 0.930 | 0.348 |
| Ensemble 2 | **0.01112 ± 0.01917** | 0.00432 (0.01078) | **0.949** | **0.964** | 0.896 | 0.913 | 0.327 |
| Ensemble 3 | 0.01184 ± 0.02109 | 0.00439 (0.01134) | 0.945 | 0.961 | 0.896 | 0.904 | 0.335 |

Table 3: Comparison of the current state-of-the-art in the field of QEI of ASL CBF Maps (Dolui et al. (2024)) with the different QEI methods presented in this study using the test data set.

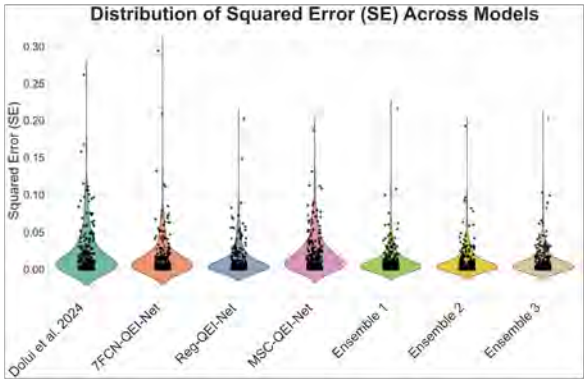
| Method | MSE ± std SE | Median of SE (IQR) | PC Coefficient | AUC | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| Dolui et al. 2024 QEI | 0.04730 ± 0.05045 | 0.02945 (0.05103) | 0.808 | 0.896 | 0.865 | 0.583 |
| 7FCN-QEI-Net | 0.02552 ± 0.03811 | 0.01256 (0.02680) | 0.844 | 0.915 | 0.757 | 0.571 |
| Reg-QEI-Net | 0.02308 ± 0.02758 | 0.01464 (0.02414) | **0.905** | **0.950** | 0.892 | 0.765 |
| MSC-QEI-Net | 0.02776 ± 0.03141 | 0.02179 (0.03967) | 0.877 | 0.909 | 0.838 | 0.625 |
| BC-Net | - | - | - | 0.946 | 0.880 | 0.705 |
| Ensemble 1 | **0.01795 ± 0.02002** | **0.00904 (0.02551)** | 0.897 | 0.946 | 0.892 | 0.750 |
| Ensemble 2 | 0.01822 ± 0.01854 | 0.01126 (0.02616) | **0.905** | 0.946 | **0.919** | 0.786 |
| Ensemble 3 | 0.01814 ± 0.01864 | 0.01153 (0.02659) | **0.905** | 0.946 | **0.919** | **0.800** |

ensembles performing slightly better than the individual algorithms, Reg-QEI-Net delivered the best performance among the individual approaches in most metrics, and its results were also comparable to those of the ensembles.

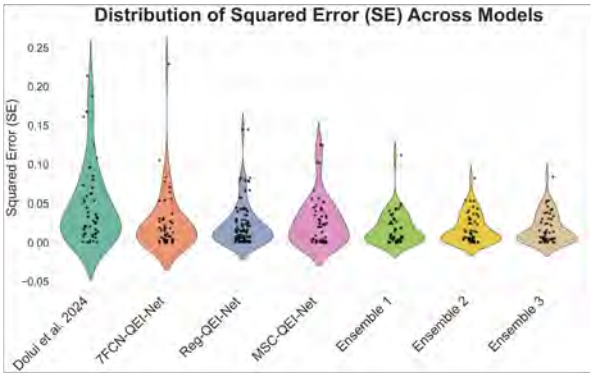
Figure 9 shows examples of the prediction using each method in 4 samples from the test set, one per rating category, in which all raters agreed with the same ratings. Each image also shows the QEI obtained using different methods, with the first entry showing the manual rating scaled in the [0,1] range. The best methods in each case, as determined by a QEI value closest to the manual rating, are shown in green. Finally, in Figure 13, we show the heatmap of the Reg-QEI-Net model, the best performer amongst the individual approaches, corresponding to various samples, each demonstrating different sources of artifacts.

### *4.2. QEI across studies*

Given that the dataset used in this study includes data from six different multisite studies, we have analyzed the performance of the presented approaches across these sources. Figure 12 shows the distribution of the QEI for each method across the different studies for both the validation and the test set. Note that the test set does not encompass all the studies. The figure also shows the color-coded manual ratings for each method. As expected, the VCID with its advanced protocol had the best QEI, while the ADNI ASL, with a relatively poor protocol and also acquired in older healthy participants and patients who are more susceptible to move, performed worst.




(a)




(b)

Figure 10: Violin plot illustrating the distribution of SE across all the methods compared in this study for (a) validation and (b) test set.

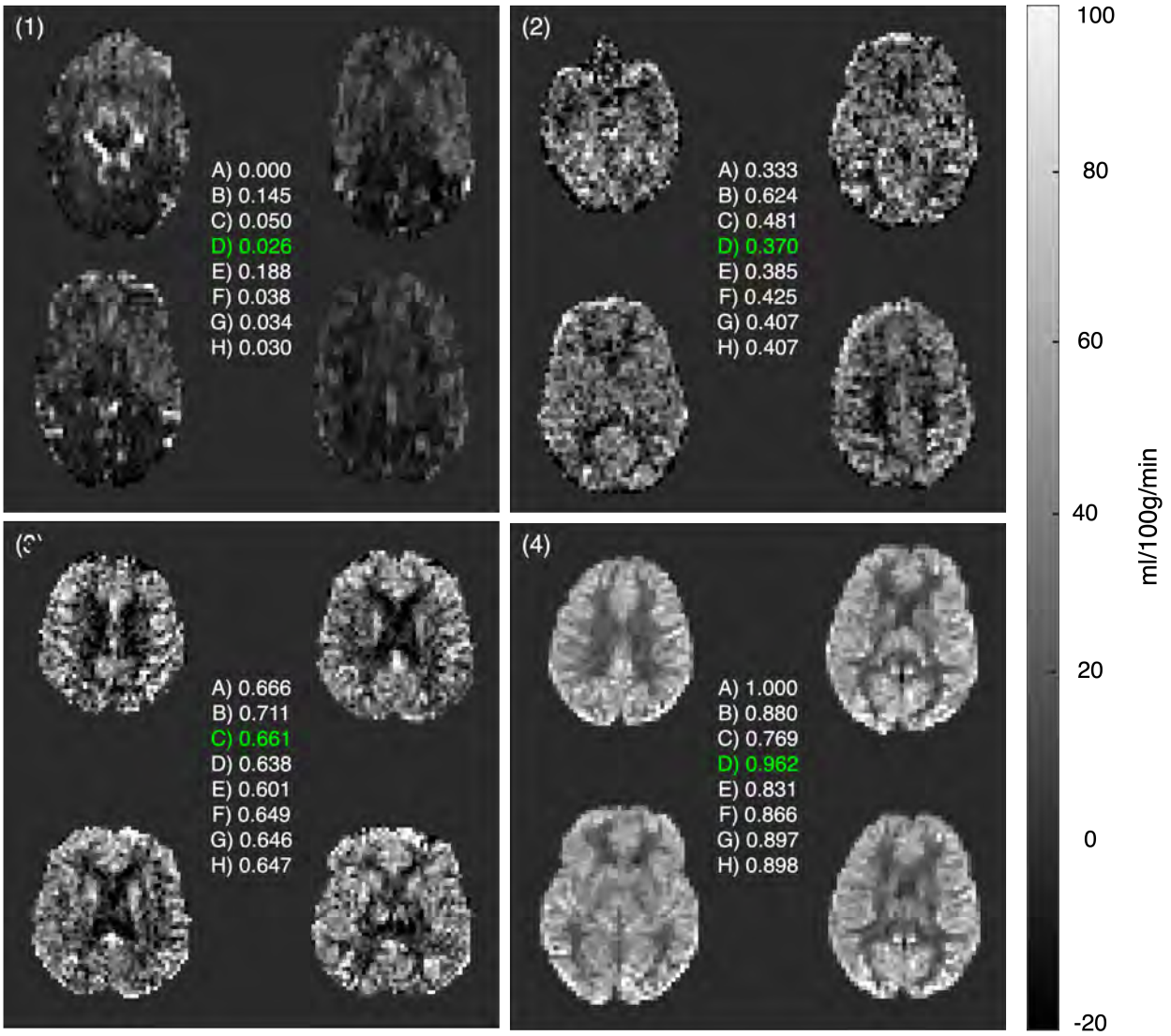



Figure 9: Example of ASL CBF Maps with (1) Unacceptable quality (Rating 1) (2) Poor quality (Rating 2) (3) Average Quality (Rating 3) and (4) Excellent Quality (Rating 4) from the test set. Each example includes the QEI prediction for each of the presented approaches. A,B,C,D,E,F,G,H correspond to the average ratings of the raters, Dolui et al. (2024), 7FCN-QEI-Net, Reg-QEI-Net, MSC-QEI-Net, Ensemble 1, Ensemble 2, and Ensemble 3, respectively.

## 5. Discussion

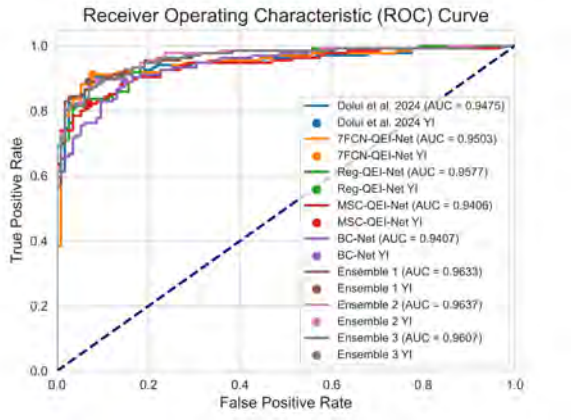
In this work, we developed several automated QEIs of ASL CBF maps by leveraging DL techniques. We improved the current state-of-the-art method (Dolui et al. (2024)) by introducing four new features and using them to train an FCN. While this method already surpassed the performance of (Dolui et al. (2024)), its limitations in the number of features and lack of automation prompted the exploration of other possibilities. To automate the feature extraction process, we developed multiple CNN approaches. These models outperformed the previous results, demonstrating the superiority of CNNs in finding better feature representations. Note that these methods only used the CBF map as input and did not require a structural image, unlike the 7-FCN-Net method, which extracts features from different tissue types. We also considered ensembles of some of the individual approaches, however, Reg-QEI-Net provided results comparable to the ensembled approaches, and is therefore our recommendation to be used clinically or in research.

### *5.1. Quality assessment methods*

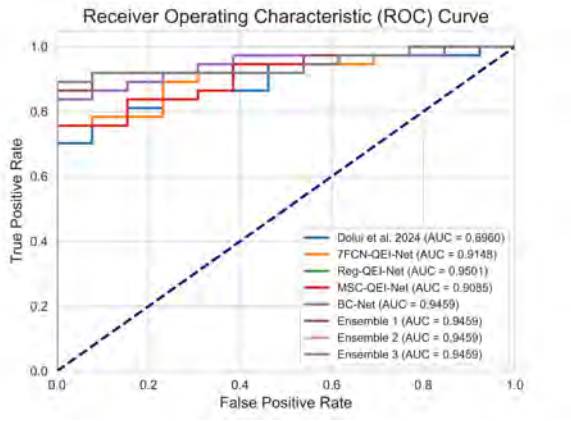
Table 2 and Table 3 present detailed comparisons of all the proposed approaches against the current state-of-the-art method (Dolui et al. (2024)) for the validation and the test sets, respectively. These results show that all the DL methods agree with the manual ratings. Specifically, the automated measures correlated better with the average ratings than the inter-rater correlation.

While all the raters are highly experienced researchers at the forefront of ASL MRI, their agreement is not perfect, highlighting the inherent difficulty and subjectivity of this task. Although not tested explicitly as a part of this study, the intra-rater agreement is also not expected to be perfect, and the agreement can be lower with raters new to the field who have limited experience with ASL CBF maps. The automated rating, being an objective measure, has the advantage of perfect reproducibility, thus increasing scientific rigor and reliability. All the DL approaches outperformed the current state-of-the-art approach (Dolui et al. (2024)). The 7FCN-QEI-Net model incorporates more features and uses a better machine learning approach to fit to the training data than (Dolui et al. (2024)), which uses a relatively naïve approach to fit each feature separately and combine them subsequently. As mentioned before, the improvements are even more pronounced with the CNN-based approaches, as showcased in Table 2 and Table 3. The MSC-QEI-Net approach performed slightly worse than Reg-QEI-Net. While that can be simply due to the nature of the problem, which inputs and outputs continuous variables, other aspects could have affected the performance of the algorithm. For example, we are currently using a categorical focal cross-entropy loss function for model training, which helps in dealing with imbalanced datasets. It might be beneficial to implement a customized weighted categorical cross-entropy loss function, where predictions are weighted according to each rater's class distribution. This approach might better address the underrepresented classes and improve overall performance.

Following the implementation of CNN-based models, we developed ensemble approaches. The goal of combining these models is based on their fundamentally different natures. For example, one model consists of a FCN, while the others are CNNs designed for completely different tasks. As a result, their performance and feature vectors vary due to their individual strengths and weaknesses. This is illustrated in Figure 12, where the networks exhibit varying levels of difficulty in predicting different rating values. For instance, in the ADNI dataset, the 7FCN-QEI-Net and Reg-QEI-Net were less successful at predicting samples rated as 2 than samples with other ratings. In contrast, MSC-QEI-Net did not encounter significant issues with samples from this rating group. Instead, this network performed less effectively when predicting samples rated as 1 and 3. The ensemble methods aim to address this by combining predictions from all models, creating a single, more robust, and more accurate final prediction. Although theoretically sound, we only found a minor improvement in performance with this approach. However, we expect further improvement when we train our models with a wider variety of ASL data from different scanners in our future work (more details in the Future Work section below).




(a)

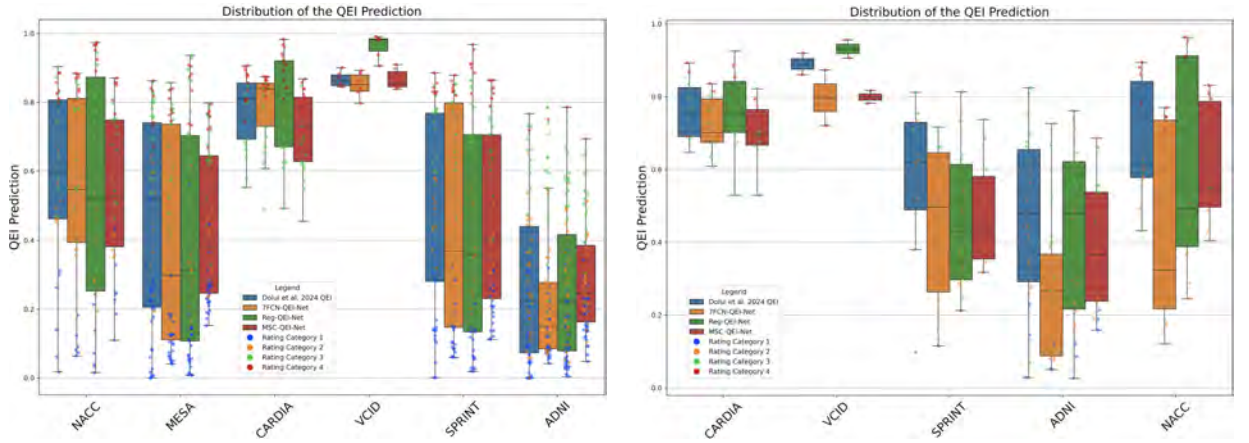



(b)

Figure 11: ROC Curve of the different approaches compared in this study corresponding to the (a) validation and the (b) test set.

### *5.2. Identifying unacceptable quality CBF Maps*

Once the QEI has been obtained from the presented methods, to exclude unacceptable CBF maps, we have presented recommendations for cut off values based on the YI, which optimizes both sensitivity and specificity. However, in research studies, the preference for higher sensitivity or higher specificity may vary depending on the specific task and the type of ASL data that is used. For example, a research study dealing with poor ASL data can drastically reduce its sample size using the optimal cutoff value. Therefore, it may be beneficial for such a study to lower the cutoff value to preserve enough data for analysis. On the other hand, a study dealing with state-of-the-art ASL data, or having a very large sample size, can use a higher QEI threshold to preserve only the ASL data with the best quality. Since the QEI produces a continuous number between 0 and 1, this provides the researcher flexibility to choose a threshold depending on the ASL data.




(a) (b)

Figure 12: The QEI values across studies for both the (a) validation and the (b) test set.

### *5.3. Interpreting the heatmaps: artifact detection*

The QEI presented in this study represents an estimate of the overall CBF map quality. A mediocre QEI value indicates that there are artifacts in the image, but it does not specify their location. The CNN-based QEI models do not provide a direct explanation for providing a low or high QEI, as the features are automatically extracted. A heatmap generated by one of their convolutional layers, however, can provide such information. This heatmap could be used for region of interest analysis. For example, for CBF maps with mediocre QEI values, the heatmaps can be used to create regions of unreliable CBF maps that can be discarded from statistical analysis. As shown in Figure 13, the higher intensities in the heatmap in artifactual CBF maps coincide with the region of artifacts. In samples free of artifacts, the network typically focuses on the GM and WM areas, where CBF is most relevant and significant. In the presented artifact-free case, the network has focused on the mentioned regions but has also shown a special interest in the right occipital lobe, identifying a potential source of artifact. This sample was originally rated as a 4 (free of artifacts) by two raters and as a 3 by the third rater. After discussing this case with the two raters who rated it as a 4, they agreed that the image might include a small amount of transit artifacts in the highlighted area. Due to their extensive expertise in ASL, the two raters knew that the protocol used for this sample was a single PLD. This protocol minimizes the transit artifact but does not eliminate it. Therefore, they concluded that this image was of very high quality (rating 4) considering the protocol used in the acquisition. The network QEI for this sample was 0.9057. This demonstrates the high correlation between the network's assessments and those of the raters, while also showcasing the network's potential ability to detect even the smallest artifacts.

### *5.4. Limitations*

This study has several limitations. First, the ASL data that was used for this study was all acquired with Siemens scanners. Hence, although the study utilizes different ASL methods, there can potentially be further variability due to differences across MRI vendor platforms that were not captured by the models and need to be studied in the future. Second, the models were trained with a very limited sample size. This study is the beginning of a 5-years project funded by the National Institutes of Health (NIH) and eventually the models will be trained with a much larger sample size, including data obtained with other scanning platforms. Third, this study did not cover all possible artifacts or disease types because of limited availability; the dataset will be expanded in the future phase of the project. Fourth, we had 3 raters who rated the images on 4 scales, which led to a limited range of unique numbers when averaged. Some raters expressed that, for certain images, they were unsure between two rating levels and would have preferred more options rather than being forced to choose one that they did not fully agree with. Therefore, it would be beneficial for this task to extend the current rating levels to a wider range, which would provide more options to the raters and a richer representation of the network's ground truth, thereby improving the model's ability to learn and perform accurately. Finally, we had only 3 raters who rated the images. Incorporating additional raters to rate the images can generalize the QEIs, as different raters might have different sensitivities to different types of artifacts.

### *5.5. Future Work*

Despite achieving state-of-the-art results, there is potential for further improvement by addressing the limitations mentioned above. First, we will aim to train the models by using a much larger dataset encompassing

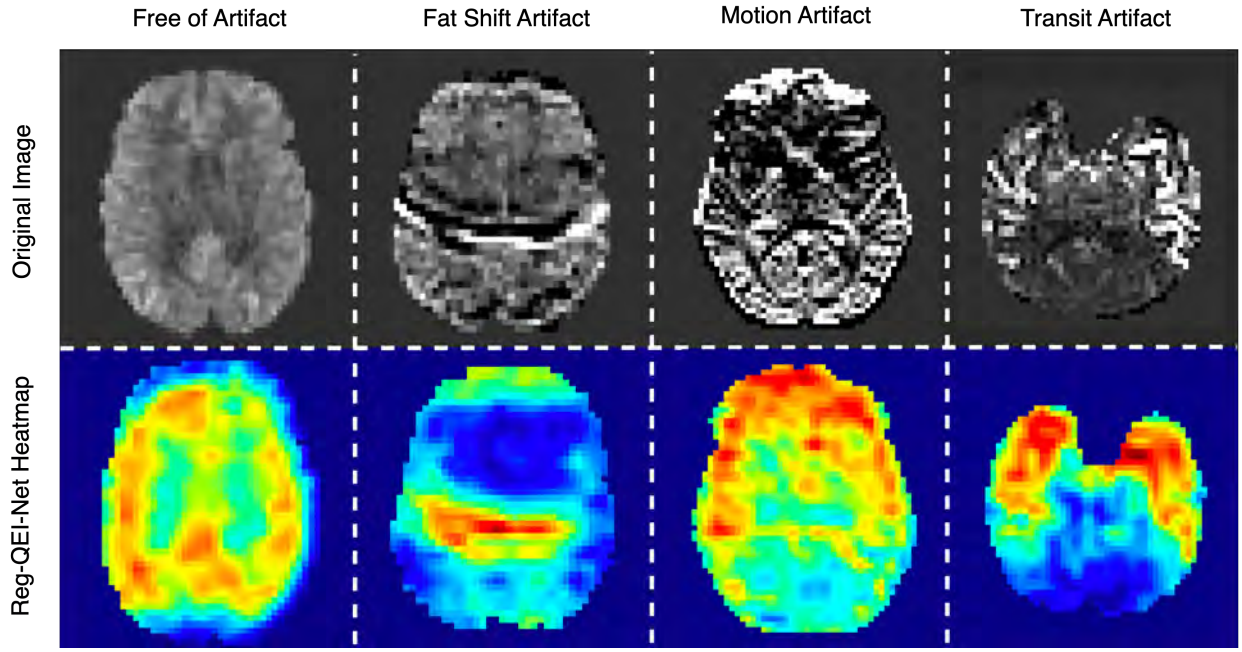



Figure 13: Example of Reg-QEI-Net heatmap visualizations applied to various samples with different sources of artifacts.

a wider variety of ASL protocols, more scanning platforms (e.g., GE and Philips), a wider type of artifact, data from patients with different diseases, and images rated by a greater number of raters. By doing so, the diversity and relevance of the training data can be increased, leading to improved network performance and robustness. Second, we will apply the QEI to actual research studies to assess improvements in statistical tests of group differences. Third, we will use heatmaps to identify regions of unreliable CBF maps and apply that to ROI analysis to assess if that improves statistical results. Lastly, although the current QEI-Net approach demonstrates high performance in assessing the quality of the CBF map, it does not give information about the source of artifacts. To address this, a CNN model aimed at classifying different sources of artifacts could be implemented, that can be used in studies to modify or correct errors in data acquisition protocols. For this approach, the heatmaps obtained from the QEI-Net architecture could serve as ROI extractors, enhancing the network's ability to focus on more meaningful areas of the brain. This improvement would not only increase the interpretability of the results but also provide valuable insights into the types of artifacts affecting the quality of the maps, ultimately contributing to better diagnostic outcomes and model transparency.

## 6. Conclusions

In this study, we designed, optimized, and validated multiple automated QEIs for ASL-derived CBF maps using DL techniques. The methods perform comparably to manual quality assessments and can rapidly provide an objective quality evaluation that can be used in research studies. These methods can also be incorporated into clinical and research scanners and provide real-time feedback to the scanner technicians that can be used to repeat the scans while the patient or study participant is still in the scanner. The automated QEI is expected to facilitate scientific rigor and reproducibility in research studies.

## Acknowledgments

I am deeply grateful to my supervisors, Dr. Sudipto Dolui and Dr. John A. Detre, for their trust, insightful feedback, and the freedom they provided throughout this project. I extend my heartfelt thanks to my family and friends for their unwavering emotional support, despite the physical distances between us. I am also sincerely appreciative of the MAIA master consortium for granting me this life-changing opportunity. Finally, this endeavor would not have been possible without the financial support of the National Institutes of Health (NIH) Grant R21AG080518.

## References

Alsop, D., Detre, J., Golay, X., et al., 2015. Recommended implementation of arterial spin-labeled perfusion mri for clinical applications: A consensus of the ismrm perfusion study group and the european consortium for asl in dementia. Magn Reson Med 73, 102–116. doi:10.1002/mrm.25197.

Austin, T.R., Nasrallah, I.M., Erus, G., Desiderio, L.M., Chen, L.Y., Greenland, P., Harding, B.N., Hughes, T.M., Jensen, P.N., Longstreth Jr, W., Post, W.S., Shea, S.J., Sitlani, C.M., Davatzikos, C., Habes, M., Bryan, R.N., Heckbert, S.R., 2024. Association of brain volumes and white matter injury with race, ethnicity, and cardiovascular risk factors: The multi-ethnic study of atherosclerosis .

Bengio, Y., Courville, A., Vincent, P., 2013. Representation learning: A review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35, 1798–1828.

Bernbaum, M., Menon, B., Fick, G., Smith, E., Goyal, M., Frayne, R., Coutts, S., 2015. Reduced blood flow in normal white matter predicts development of leukoaraiosis. Journal of Cerebral Blood Flow and Metabolism 35, 1610–1615. doi:10.1038/jcbfm.2015.92.

Binnewijzend, M., Kuijer, J., Benedictus, M., van der Flier, W., Wink, A., Wattjes, M., van Berckel, B., Scheltens, P., Barkhof, F., 2013. Cerebral blood flow measured with 3d pseudocontinuous arterial spin-labeling mr imaging in alzheimer disease and mild cognitive impairment: a marker for disease severity. Radiology 267, 221–230. doi:10.1148/radiol.12120928.

Buxton, R., Frank, L., Wong, E., Siewert, B., Warach, S., Edelman, R., 1998. A general kinetic model for quantitative perfusion imaging with arterial spin labeling. Magnetic Resonance in Medicine 40, 383–396. doi:10.1002/mrm.1910400308.

De La Torre, J., 2013. Vascular risk factors: A ticking time bomb to alzheimer's disease. American Journal of Alzheimer's Disease and other Dementias 28, 551–559. doi:10.1177/1533317513494457.

Deng, L., Dong, Y., 2014. Deep learning: Methods and applications. Found Trends® Signal Process 7, 197–387.

Detre, J., Leigh, J., Williams, D., Koretsky, A., 1992. Perfusion imaging. Magn Reson Med 23, 37–45. doi:10.1002/mrm.1910230106.

Dolui, S., Detre, J., Gaussoin, S., Herrick, J., Wang, D., Tamura, M., Cho, M., Haley, W., Launer, L., Punzi, H., Rastogi, A., Still, C., Weiner, D., Wright, J.J., Williamson, J., Wright, C., Bryan, R., Bress, A., Pajewski, N., Nasrallah, I., 2022. Association of intensive vs standard blood pressure control with cerebral blood flow: Secondary analysis of the sprint mind randomized clinical trial. JAMA neurology 79, 380–389. doi:10.1001/jamaneurol.2022.0074.

Dolui, S., Li, Z., Nasrallah, I., Detre, J., Wolk, D., 2020. Arterial spin labeling versus (18)f-fdg-pet to identify mild cognitive impairment. NeuroImage Clinical 25, 102146. doi:10.1016/j.nicl.2019.102146.

Dolui, S., Tisdall, D., Vidorreta, M., et al., 2019. Characterizing a perfusion-based periventricular small vessel region of interest. NeuroImage Clinical 23, 101897.

Dolui, S., Vidorreta, M., Wang, Z., Nasrallah, I., Alavi, A., Wolk, D., Detre, J., 2017a. Comparison of pasl, pcasl, and background-suppressed 3d pcasl in mild cognitive impairment. Human Brain Mapping 38, 5260–5273. doi:10.1002/hbm.23732.

Dolui, S., Wang, Z., Shinohara, R., Wolk, D., Detre, J., I, A.D.N., 2017b. Structural correlation-based outlier rejection (score) algorithm for arterial spin labeling time series. Journal of Magnetic Resonance Imaging 45, 1786–1797. doi:10.1002/jmri.25436.

Dolui, S., Wang, Z., Wang, D.J., et al., 2016. Comparison of non-invasive mri measurements of cerebral blood flow in a large multisite cohort. Journal of Cerebral Blood Flow & Metabolism 36, 1244–1256. doi:10.1177/0271678X16646124.

Dolui, S., Wang, Z., Wolf, R., Nabavizadeh, A., Xie, L., Tosun, D., Nasrallah, I., Wolk, D., Detre, J., I, A.D.N., 2024. Automated quality evaluation index for arterial spin labeling derived cerebral blood flow maps. Journal of magnetic resonance imaging: JMRI doi:10.1002/jmri.29308.

Ewing, J., Cao, Y., Knight, R., Fenstermacher, J., 2005. Arterial spin labeling: validity testing and comparison studies. Journal of Magnetic Resonance Imaging 22, 737–740. doi:10.1002/jmri.20451.

Fallatah, S., Pizzini, F., Gómez-Anson, B., Magerkurth, J., Vita, E., Bisdas, S., Jäger, H., Mutsaerts, H., Golay, X., 2018. A visual quality control scale for clinical arterial spin labeling images. European Radiology Experimental 2, 1–8.

Fazlollahi, A., Calamante, F., Liang, X., Bourgeat, P., Raniga, P., Dore, V., Fripp, J., Ames, D., Masters, C., Rowe, C., Connelly, A., Villemagne, V., Salvado, O., B, A.I., G, L.R., 2020. Increased cerebral blood flow with increased amyloid burden in the preclinical phase of alzheimer's disease. Journal of Magnetic Resonance Imaging 51, 505–513. doi:10.1002/jmri.26810.

Fernandez-Seara, M., Wang, Z., Wang, J., Rao, H., Guenther, M., Feinberg, D., Detre, J., 2005. Continuous arterial spin labeling perfusion measurements using single shot 3d grase at 3 t. Magn Reson Med 54, 1241–1247.

Friston, K., Williams, S., Howard, R., Frackowiak, R., Turner, R., 1996. Movement-related effects in fmri time-series. Magn Reson Med 35, 346–355.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks, in: Teh, Y.W., Titterington, M. (Eds.), Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, PMLR, Chia Laguna Resort, Sardinia, Italy. pp. 249–256. URL: https://proceedings.mlr.press/v9/glorot10a.html.

Gong, E., Guo, J., Liu, J., Fan, A., Pauly, J., Zaharchuk, G., 2020. Deep learning and multi-contrast based denoising for low-snr arterial spin labeling (asl) mri, in: Medical Imaging 2020: Image Processing, SPIE. p. 113131K. doi:10.1117/12.2549765.

Gregg, N., Kim, A., Gurol, M., et al., 2015. Incidental cerebral microbleeds and cerebral blood flow in elderly individuals. JAMA Neurology 72, 1021–1028. doi:10.1001/jamaneurol.2015.1359.

He, S., Feng, Y., Grant, P., Ou, Y., 2022. Deep relation learning for regression and its application to brain age estimation. IEEE Transactions on Medical Imaging 41, 2304–2317. doi:10.1109/TMI.2022.3161739.

Heijtel, D., Mutsaerts, H., Bakker, E., Schober, P., Stevens, M., Petersen, E., van Berckel, B., Majoie, C., Booij, J., van Osch, M., Vanbavel, E., Boellaard, R., Lammertsma, A., Nederveen, A., 2014. Accuracy and precision of pseudo-continuous arterial spin labeling perfusion during baseline and hypercapnia: a head-to-head comparison with (1)(5)o h(2)o positron emission tomography. NeuroImage 92, 182–192. doi:10.1016/j.neuroimage.2014.02.011.

Herscovitch, P., Markham, J., Raichle, M., 1983. Brain blood flow measured with intravenous h2(15)o. i. theory and error analysis. Journal of nuclear medicine: official publication, Society of Nuclear Medicine 24, 782–789.

Ito, H., Inoue, K., Goto, R., Kinomura, S., Taki, Y., Okada, K., Sato, K., Sato, T., Kanno, I., Fukuda, H., 2006. Database of normal human cerebral blood flow measured by spect: I. comparison between i-123-imp, tc-99m-hmpao, and tc-99m-ecd as referred with o-15 labeled water pet and voxel-based morphometry. Annals of nuclear medicine 20, 131–138.

Iturria-Medina, Y., Sotero, R., Toussaint, P., Mateos-Perez, J., Evans, A., I, A.D.N., 2016. Early role of vascular dysregulation on late-onset alzheimer's disease based on multifactorial data-driven analysis. Nat Commun 7, 11934. doi:10.1038/ncomms11934.

Jain, V., Langham, M., Wehrli, F., 2010. Mri estimation of global brain oxygen consumption rate. Journal of cerebral blood flow and metabolism: official journal of the International Society of Cerebral Blood Flow and Metabolism 30, 1598–1607. doi:10.1038/jcbfm.2010.49.

Kety, S., Schmidt, C., 1945. The determination of cerebral blood flow in man by the use of nitrous oxide in low concentrations. Am J Physiol 143, 53–66.

Kim, K.H., Choi, S.H., Park, S.H., 2018. Improving arterial spin labeling by using deep learning. Radiology 287, 658–666. doi:10.1148/radiol.2017171154.

Koenig, M., Klotz, E., Luka, B., Venderink, D., Spittler, J., Heuser, L., 1998. Perfusion ct of the brain: diagnostic approach for early detection of ischemic stroke. Radiology 209, 85–93. doi:10.1148/radiology.209.1.9769817.

Krizhevsky, A., Sutskever, I., Hinton, G., 2012. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst .

Lassen, N., Henriksen, L., Paulson, O., 1981. Regional cerebral blood flow in stroke by 133xenon inhalation and emission tomography. Stroke; a journal of cerebral circulation 12, 284–288.

Lecun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

Leonardsen, E., Peng, H., Kaufmann, T., Agartz, I., Andreassen, O., Celius, E., Espeseth, T., Harbo, H., Høgestøl, E., Lange, A., Marquand, A., Vidal-Piñeiro, D., Roe, J., Selbæk, G., Sørensen, Y., Smith, S., Westlye, L., Wolfers, T., Wang, Y., 2022. Deep neural networks learn general and clinically relevant representations of the ageing brain. NeuroImage 256, 119210. doi:10.1016/j.neuroimage.2022.119210.

Li, Y., Dolui, S., Xie, D., Wang, Z., Initiative, A., 2018. Priors-guided slice-wise adaptive outlier cleaning for arterial spin labeling perfusion mri. Journal of Neuroscience Methods 307, 248–253. doi:10.1016/j.jneumeth.2018.06.007.

Li, Y., Liu, P., Li, Y., et al., 2019. Asl-mricloud: An online tool for the processing of asl mri data. NMR in Biomedicine 32, e4051. doi:10.1002/nbm.4051.

Litjens, G., Kooi, T., Bejnordi, B., Setio, A., Ciompi, F., Ghafoorian, M., van der Laak, J., van Ginneken, B., Sánchez, C., 2017. A survey on deep learning in medical image analysis. Med Image Anal 42, 60–88.

Maleki, N., Dai, W., Alsop, D., 2012. Optimization of background suppression for arterial spin labeling perfusion imaging. Magnetic Resonance Materials in Physics, Biology and Medicine 25, 127–133.

Moonen, J., Nasrallah, I., Detre, J., Dolui, S., Erus, G., Davatzikos, C., Meirelles, O., Bryan, R., Launer, L., 2020. Race and sex differences in midlife changes in cerebral volume and perfusion. Alzheimer's Dement 16, 1–2.

Peng, H., Gong, W., Beckmann, C., Vedaldi, A., Smith, S., 2021. Accurate brain age prediction with lightweight deep neural networks. Medical Image Analysis 68, 101871. doi:10.1016/j.media.2020.101871.

Plis, S., Hjelm, D., Salakhutdinov, R., Allen, E., Bockholt, H., Long, J., Johnson, H., Paulsen, J., Turner, J., Calhoun, V., 2014. Deep learning for neuroimaging: a validation study. Frontiers in Neuroscience 8, 229. doi:10.3389/fnins.2014.00229.

Power, J., Barnes, K., Snyder, A., Schlaggar, B., Petersen, S., 2012. Spurious but systematic correlations in functional connectivity mri networks arise from subject motion. Neuroimage 59, 2142–2154.

Qin, Q., Alsop, D., Bolar, D., Hernandez-Garcia, L., Meakin, J., Liu, D., Nayak, K., Schmid, S., van Osch, M., Wong, E., Woods, J., Zaharchuk, G., Zhao, M., Zun, Z., Guo, J., Group, I., 2022. Velocity-selective arterial spin labeling perfusion mri: A review of the state of the art and recommendations for clinical implementation. Magnetic Resonance in Medicine 88, 1528–1547.

Rempp, K., Brix, G., Wenz, F., Becker, C., Guckel, F., Lorenz, W., 1994. Quantification of regional cerebral blood flow and volume with dynamic susceptibility contrast-enhanced mr imaging. Radiology 193, 637–641. doi:10.1148/radiology.193.3.7972800.

Ruopp, M., Perkins, N., Whitcomb, B., Schisterman, E., 2008. Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. Biometrical Journal 50, 419–430. doi:10.1002/bimj.200710415.

Sadaghiani, S., Tackett, W., Tisdall, M., Detre, J., Dolui, S., 2023. Reliability of periventricular white matter cerebral blood flow using different asl protocols. Proceedings on CD-ROM - International Society for Magnetic Resonance in Medicine. Scientific Meeting and Exhibition/Proceedings of the International Society for Magnetic Resonance in Medicine, Scientific Meeting and Exhibition doi:10.58530/2022/4875.

Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV) , 618–626.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 doi:10.48550/arXiv.1409.1556.

Taso, M., Aramendía-Vidaurreta, V., Englund, E., Francis, S., Franklin, S., Madhuranthakam, A., Martirosian, P., Nayak, K., Qin, Q., Shao, X., Thomas, D., Zun, Z., Fernández-Seara, M., Group, I., 2023. Update on state-of-the-art for arterial spin labeling (asl) human perfusion imaging outside of the brain. Magnetic Resonance in Medicine 89, 1754–1776. doi:10.1002/mrm.29609.

Wang, Z., Das, S.R., Xie, S.X., et al., 2013. Arterial spin labeled mri in prodromal alzheimer's disease: A multi-site study. NeuroImage: Clinical 2, 630–636. doi:10.1016/j.nicl.2013.04.014.

Williams, D., Detre, J., Leigh, J., Koretsky, A., 1992. Magnetic resonance imaging of perfusion using spin inversion of arterial water. Proc Natl Acad Sci U S A 89, 212–216. doi:10.1073/pnas.89.1.212.

Wolk, D., Detre, J., 2012. Arterial spin labeling mri: an emerging biomarker for alzheimer's disease and other neurodegenerative conditions. Current opinion in neurology 25, 421–428. doi:10.1097/WCO.0b013e328354ff0a.

Woods, J., Achten, E., Asllani, I., Bolar, D., Dai, W., Detre, J., Fan, A., Fernandez-Seara, M., Golay, X., Gunther, M., Guo, J., Hernandez-Garcia, L., Ho, M., Juttukonda, M., Lu, H., MacIntosh, B., Madhuranthakam, A., Mutsaerts, H., Okell, T., Parkes, L., Pinter, N., Pinto, J., Qin, Q., Smits, M., Suzuki, Y., Thomas, D., Van Osch, M., Wang, D., Warnert, E., Zaharchuk, G., Zelaya, F., Zhao, M., Chappell, M., Group, I., 2024. Recommendations for quantitative cerebral perfusion mri using multi-timepoint arterial spin labeling: Acquisition, quantification, and clinical applications. Magnetic Resonance in Medicine 92, 469–495.

Xie, D., Li, Y., Yang, H., et al., 2020. Denoising arterial spin labeling perfusion mri with deep machine learning. Magnetic Resonance Imaging 68, 95–105. doi:10.1016/j.mri.2020.01.005.

Ye, F., Berman, K., Ellmore, T., Esposito, G., van Horn, J., Yang, Y., Duyn, J., Smith, A., Frank, J., Weinberger, D., McLaughlin, A., 2000a. H(2)(15)o pet validation of steady-state arterial spin tagging cerebral blood flow measurements in humans. Magnetic Resonance in Medicine 44, 450–456.

Ye, F., Frank, J., Weinberger, D., McLaughlin, A., 2000b. Noise reduction in 3d perfusion imaging by attenuating the static signal in arterial spin tagging (assist). Magn Reson Med 44, 92–100.

Yonas, H., Darby, J., Marks, E., Durham, S., Maxwell, C., 1991. Cbf measured by xe-ct: approach to analysis and normal values. Journal of cerebral blood flow and metabolism: official journal of the International Society of Cerebral Blood Flow and Metabolism 11, 716–725. doi:10.1038/jcbfm.1991.128.

Zhang, K., Herzog, H., Mauler, J., Filss, C., Okell, T., Kops, E., Tellmann, L., Fischer, T., Brocke, B., Sturm, W., Coenen, H., Shah, N., 2014. Comparison of cerebral blood flow acquired by simultaneous [15o]water positron emission tomography and arterial spin labeling magnetic resonance imaging. Journal of cerebral blood flow and metabolism: official journal of the International Society of Cerebral Blood Flow and Metabolism 34, 1373–1380. doi:10.1038/jcbfm.2014.92.

Zhang, L., Xie, D., Li, Y., et al., 2022. Improving sensitivity of arterial spin labeling perfusion mri in alzheimer's disease using transfer learning of deep learning-based asl denoising. Journal of Magnetic Resonance Imaging 55, 1710–1722. doi:10.1002/jmri.27984.

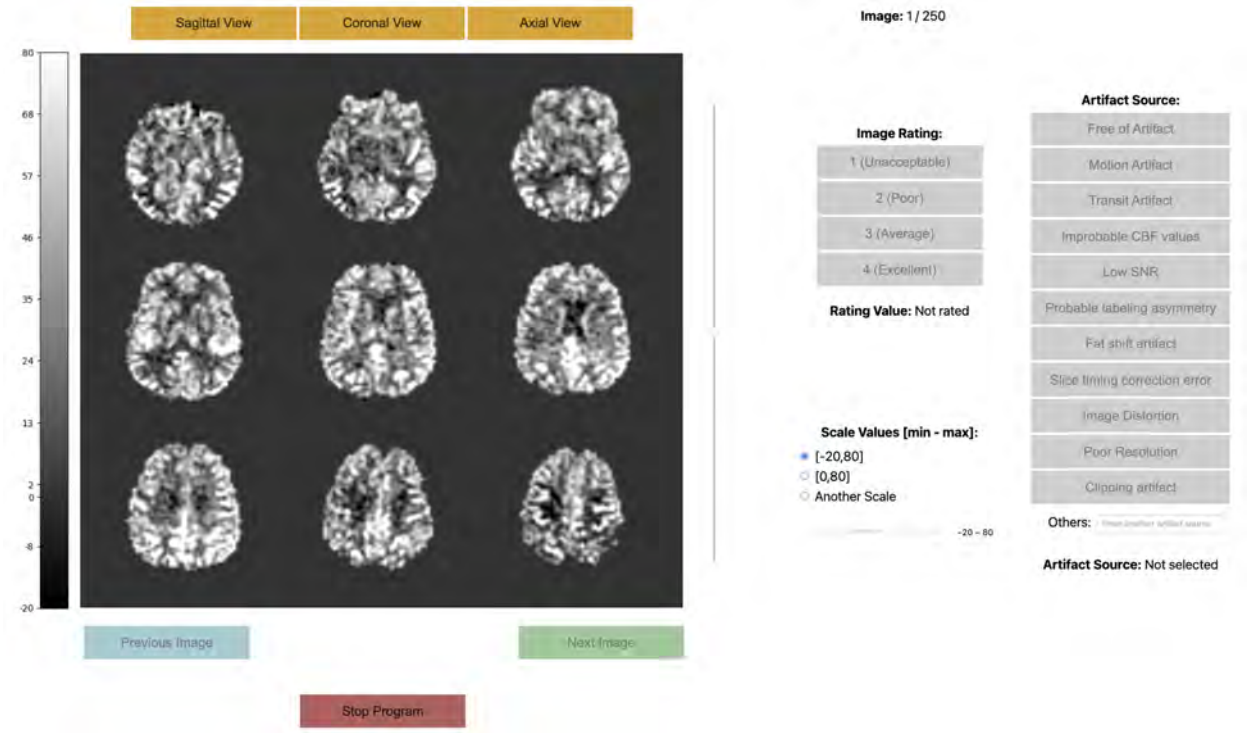



Figure 14: Example of the ASL CBF rating tool.

## Appendix A: ASL CBF Rating Tool

Here, we detail the functionality and features of the web-based ASL CBF rating tool developed to simplify the rating task. This tool is a Python notebook designed to be used with Google Collaboratory, thereby eliminating the need to install software or sensitive data on the user's computer. When the script is initiated, it automatically downloads the dataset from Dropbox to the user's Google Drive. It also generates an Excel file where the ratings are stored.

Some of the tool's characteristics are:

- **Pause and Resume Capability:** The tool allows for pausing and resuming at any point. It automatically checks the Excel file to determine the last image rating, ensuring a seamless continuation of the task.
- **Artifact Documentation:** As part of an upcoming study on classifying imaging artifacts, raters are required to identify and document the sources of any artifacts observed.
- **Intensity Clipping:** To modify image contrast, intensity clipping is employed with default parameters set to [-20, 80].
- **Comprehensive Visualization:** The tool provides multiple views (axial, sagittal, and coronal) of each image. To enable the user to rate the image, all image views must be observed.
- **3D Navigation:** A sliding bar is included to navigate through all slices of the 3D images.

Once all images have been rated, the Excel file is automatically downloaded to the user's computer. This tool is licensed freely and is accessible via the following link: https://github.com/xavibeltranurbano/ASL-CBF-Rating-Tool

Medical Imaging and Applications

Master Thesis, June 2024

# Deep learning approaches for detecting Large Vessel Occlusion in CTA images of stroke patients

Amina Bouzid, Valeriia Abramova, Arnau Oliver, Xavier Llado

*VICOROB Research Group, Girona, Spain*

**Abstract**

Large vessel occlusions (LVOs) are a major subtype of acute ischemic stroke. They result from total or partial blockages in the brain's main arteries, which significantly hinder blood flow and cause rapid tissue death. The timely detection of LVOs is critical for improving patient outcomes. Endovascular thrombectomy (EVT), the most effective treatment, works best within 6 hours after symptom start. This study investigates the potential of deep learning for rapid and accurate LVO detection using Computed Tomography Angiography (CTA) images. We evaluated two deep learning approaches, nnDetection and 3D CNN, on a dataset of 124 CTA scans from stroke patients with LVOs acquired at Hospital Dr. Josep Trueta (Girona, Spain). The nnDetection framework achieved promising results in accurately localizing occlusions, particularly in the Anterior Circulation system, with a high sensitivity of around 90% and a low false positive per image (FPpI) rate at an optimal threshold. Encouragingly, feedback from collaborating neurologists suggests the model's detection capabilities surpass those of currently used commercial software, highlighting its potential clinical value. The 3D CNN model, designed for a different LVO detection approach, encountered challenges with high FPpI and computational demands. The study proposes several strategies to improve its performance, including utilizing dual classifiers, adjusting loss functions, and data augmentation. These findings highlight the promise of the nnDetection framework for accurate LVO detection in CTAs. Future efforts will focus on optimizing both models for improved performance and reduced computational requirements. This includes exploring advanced training techniques, expanding the data to encompass a wider range of occlusion types and anatomical variations, and conducting generalizability studies using data from different hospitals and scanner types. Ultimately, these advancements aim to develop more efficient and accurate deep learning tools for early stroke detection and treatment, potentially leading to improved patient outcomes.

*Keywords:* 3D object detection, medical detection, LVO, Ischemic Stroke, CTA, Deep learning, nnDetection

## 1. Introduction

Acute Ischemic Stroke (AIS) is a critical medical condition characterized by the blockage of a cerebral artery, resulting in the sudden interruption of the blood supply and subsequent damage to distinct cerebral regions. This blockage can be caused by a blood clot or plaque buildup in the arteries that supply nutrients to the brain tissue. There are two main types of stroke: ischemic, which accounts for over 85% of cases, and hemorrhagic. Annually, approximately 15 million people worldwide experience a stroke, with prevalence increasing with age, according to the World Health Organization and Tsao et al. (2023). Of these, around 5 million die, and another 5 million are left permanently disabled.

A specific type of AIS, known as large vessel occlusions (LVO), refers to the complete or partial blockage of one of the brain's major arteries. LVOs, which affect both anterior and posterior circulation, are responsible for approximately 46% of acute ischemic strokes. About two-thirds of these cases occur in the anterior circulation, primarily affecting the Internal Carotid Artery (ICA) and the Middle Cerebral Artery (M1). The remaining cases occurs in the posterior circulation, including the Posterior Cerebral Artery (PCA), Basilar Artery, and Vertebral Artery (see Fig. 1).

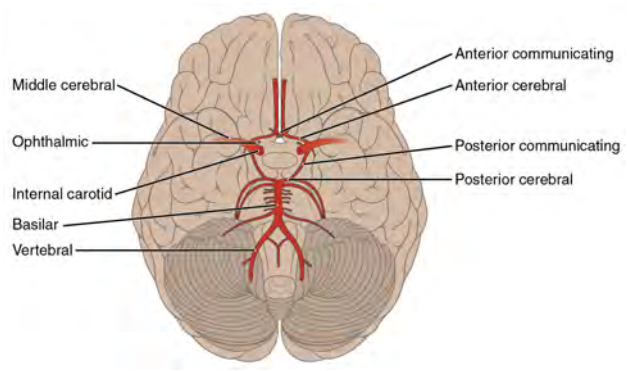



Figure 1: Brain arterial circulation. Taken from JoeNiekroFoundation (2017)

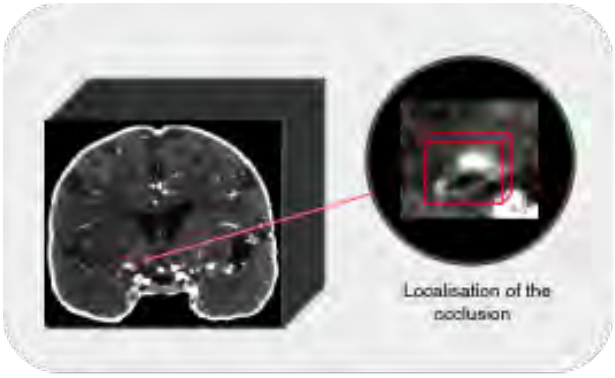



Figure 2: Example of how LVO manifests in CTA images. The red bounding box indicates the localized area of occlusion.

Another type of occlusion that can occur is called tandem, which occurs in less than 10% of cases and refers to an occlusion in more than one artery: a large blood vessel, such as the ICA, or an intracranial artery (Sweid, 2019). The damage provoked by LVOs depends mainly on the location of the occlusion and on the time of blocking.

Patients suffering from AIS related to an LVO experience the highest levels of morbidity and death, and they have the lowest probability of obtaining arterial recanalization through a clot-disolving medication called intravenous thrombolysis, as reported in the work of Martins-Filho et al. (2019). However, recent trials have presented strong evidence supporting the effectiveness of endovascular mechanical thrombectomy in treating such cases. Consequently, it is crucial to identify and transfer LVO patients to specialized stroke centers as soon as possible in order to enable rapid detection and offer suitable endovascular treatment. Effective vascular imaging techniques that can quickly detect LVO are essential, as mechanical thrombectomy is a time-sensitive procedure (Mayer, 2020).

Several imaging modalities are utilized in stroke diagnosis, with the primary ones being Non-Contrast Computed Tomography (NCCT), Computed Tomography Angiography (CTA) and Computed Tomography Perfusion (CTP). NCCT, which involves taking X-ray images of the brain without contrast agents, provides information about the presence of bleeding, tumors, or other abnormalities. This modality is commonly used as the initial imaging technique for stroke patients since it helps rule out conditions that can mimic a stroke. However, NCCT is not very sensitive to detect early signs of stroke or small infarctions. CTA, which visualizes the blood vessels in the brain using a contrast dye, is a reliable method for detecting LVOs and assessing if a patient is a good candidate for a mechanical thrombectomy (Shafaat, 2023). Fig. 2 illustrates how LVOs are visible in Computed Tomography Angiography (CTA) images, with the red bounding box indicating the localized area of occlusion.

Moreover, CTP is primarily used to evaluate the passage of blood through the tissues using a series of rapid CT scans, helping clinicians evaluate the tissue viability and identify areas of reduced blood flow.

As previously stated, endovascular thrombectomy is the established treatment for patients exhibiting stroke symptoms within a 24-hour window, as its effectiveness diminishes beyond this timeframe. Time is crucial in such cases. The treatment's goal is to reestablish 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 for this treatment is considered to be within 6 hours after the onset of symptoms. Every 30-minute delay reduces the likelihood of a good outcome by 11% (Sweid, 2019). This highlights the importance of reducing stroke care timing, including the implementation of automated tools for LVO detection.

Despite the efforts, there is still a need to standardize stroke detection and triage, which are time-sensitive processes (Murray, 2020). The implementation of automated imaging-based tools for detecting LVO has shown improvements in the timing of endovascular thrombectomy decision-making, ultimately resulting in enhanced clinical outcomes. Several commercial solutions have addressed this application. iSchevaView RapidAI[1], MethinksLVO[2], Brainomix AI, and StrokeViewer are among the noteworthy examples mentioned by Murray (2020) and Chavva (2022), demonstrating their utilization in the identification of LVOs, diagnosis of ischemic or hemorrhagic strokes, and the assessment of potentially salvageable tissue. But there are still issues with standardizing these software solutions' validation procedures and integrating them into various data streams, which makes it difficult to compare them effectively with new algorithms (Chavva, 2022).

[1]https://www.ischeva.ai/ischevaview-rapidaid/
[2]https://www.methinksapp.com/

Recent advancements in deep learning (DL) models have demonstrated significant promise in LVO detection. For instance, (Brugnara, 2023) developed an artificial neural network (ANN) capable of automated detection of abnormal vessel findings without any a-priori restrictions and in <2 minutes, and demonstrated high sensitivity (≥87%) and negative predictive value (≥93%). This study demonstrates the potential of deep learning models to improve LVO detection accuracy, potentially leading to better outcomes for patients with acute ischemic stroke. Future research directions include exploring the use of DL models for LVO detection in other imaging modalities and developing models that can not only detect LVOs but also predict their severity and assess collateral circulation, thereby significantly improving patient management in acute ischemic stroke.

### *1.1. Goals of The Master's Thesis*

The primary objective of this master's thesis is to develop and evaluate deep learning approaches for the automatic detection of Large Vessel Occlusions (LVOs) in stroke patients using Computed Tomography Angiography (CTA) images. Precise localization of LVOs is crucial for improving patient outcomes and ensuring prompt and effective treatment. To achieve this goal, we will explore two different 3D deep learning detection methods. Firstly, we will investigate the use of the nnDetection framework, which employs the Retina U-Net model. The framework is well-regarded for its robustness and high performance in medical image identification tasks. Secondly, we will utilize a 3D Convolutional Neural Network (3D CNN) approach, as described by Liao and Song. (2019). We aim to compare and assess the effectiveness of these two methods in the 3D detection of LVOs. Our evaluation will be based on a dataset consisting of 124 CTA scans annotated for LVOs. We will adopt a cross-validation strategy to ensure the reliability and robustness of our performance evaluation. TThe evaluation metrics will include sensitivity, false positive rate, and computational efficiency.

## 2. State of the art

While there are several ways that AI can be used in clinical practice to manage stroke, in this section, we will focus mainly on stroke management research that deals with the classification and detection of thrombi causing LVOs, as well as general object detection approaches.

### *2.1. LVO Classification and Detection*

Research in stroke management has significantly focused on the classification and detection of thrombi causing large vessel occlusions (LVOs). Recent initiatives, such as the IACTA-EST 2023 challenge[3], have amplified research efforts in this field, particularly in the application of CTA for stroke treatment via endovascular therapy. Deep learning techniques have become increasingly prominent in this domain, with studies showing encouraging results in thrombus detection and classification using CTA images. For instance, Stib (2020) proposed 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 and achieving high sensitivity and specificity by leveraging three phases of CT angiographies. Meijs (2020) introduced a 4D-CTA method for detecting intracranial anterior circulation occlusions with high sensitivity and specificity, though it lacks direct occlusion localization.

Barman et al. (2019) developed DeepSymNet, a CNN architecture leveraging brain symmetry for 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 on this work, Czap (2022) introduced an enhanced version of the same algorithm by adding symmetrical and unsymmetrical pathways. The latest version, DeepSymNet v3 (Giancardo et al., 2023), inputs both hemispheres separately into a network of 3D VGG blocks with shared weights between data paths. This research aims to create a complete segmentation of the stroke core using deep learning architecture. Lal-Trehan Estrada et al. (2024) used DeepSymNet v3 to detect LVOs in acute ischemic stroke patients using brain computed tomography angiography (CTA). The researchers compared strategies to enhance the network's focus on the vasculature. The results demonstrated that the proposed strategies improved LVO detection, achieving an AUC of 0.931 when combining brain CTA and 3D vasculature.

Most studies in LVO detection use existing software. Mojtahedi et al. (2022) employed StrokeViewer LVO software for bounding box localization and dual-modality U-Net for segmentation. Bruggeman et al. (2022) assessed Nico.lab's LVO algorithm, noting occlusion detection challenges and false positive rates. Other researchers use a region of interest (ROI) or bounding box provided by experts, using non-contrast CT (NCCT) images or CT angiograms (CTAs). This is likely because detection tasks are often associated with segmentation of the thrombus, causing the LVO. Consequently, the primary objective of most research in this area is to segment the potential blood clot or thrombus. For example, Lucas and Heinrich (2019) employed a U-Net architecture with ROIs defined as the union of MCA and ICA clot segmentations with a 5-voxel margin. Tolhuisen et al. (2020) proposed patch-based CNNs

[3]https://lgiancauth.github.io/iacta-est-2023/

to detect LVOs based on brain asymmetry and hyperdense artery sign (HAS), followed by voxel-wise segmentation on patches identified as containing thrombi. However, their results exhibited limited volumetric and spatial agreement.

Despite comprehensive literature, detailed descriptions of three-dimensional automatic occlusion detection models using solely CTA images for both anterior and posterior circulation remain limited. Notable exceptions include Brugnara (2023) and Bagcilar (2023), who utilized the adaptable Detection Framework (Baumgartner, 2021). This framework is an adaptive self-supervising method applicable to diverse medical detection problems, and provides a standardized interface for different datasets. Its efficacy has been demonstrated in other medical imaging tasks, highlighting its potential for LVO detection. For example, Bagcilar (2023) utilized it and achieved an AUC of 0.97 for automated LVO detection and collateral scoring on CTA scans using a multi-task 3D object detection approach. the nnDetection model was trained on a large-scale, multi-center, heterogeneous dataset (e.g., various centers, scanner vendors, CTA acquisition protocols, contrast phases, etc.), achieved an accuracy exceeding 98% in identifying LVO on independent external test data. Moreover, the nnDetection model demonstrated strong agreement in assigning collateral scores, exhibiting performance comparable to or surpassing individual radiologists' reliability when considering the radiologists' consensus as ground truth.

In summary, ongoing research in LVO classification and detection underscores the critical role of AI in stroke management, with notable progress in leveraging deep learning techniques and self-supervised methods. However, challenges persist in standardization, integration, and algorithm validation, highlighting the need for continued research and collaboration in this domain.

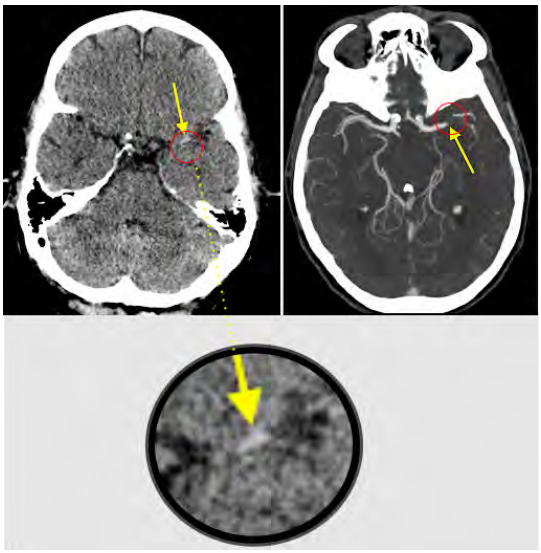
### *2.2. General Object Detection*

General object detection aims to localize and classify objects of various categories in an image. This task is typically accomplished using one-stage or two-stage detectors. One-stage detectors consider object detection as a regression problem, using a unified framework to estimate class probabilities and bounding box coordinates, enabling faster inference. YOLO (Redmon et al., 2016) and SSD (Liu and Berg, 2016) are examples of one-stage detectors. Two-stage detectors utilize a region proposal network (RPN) to generate regions of interest (ROI). A deep neural network is then applied to each proposal for classification. Examples include Faster R-CNN (Ren et al., 2015) and Mask R-CNN, which achieve higher accuracy but slower inference.

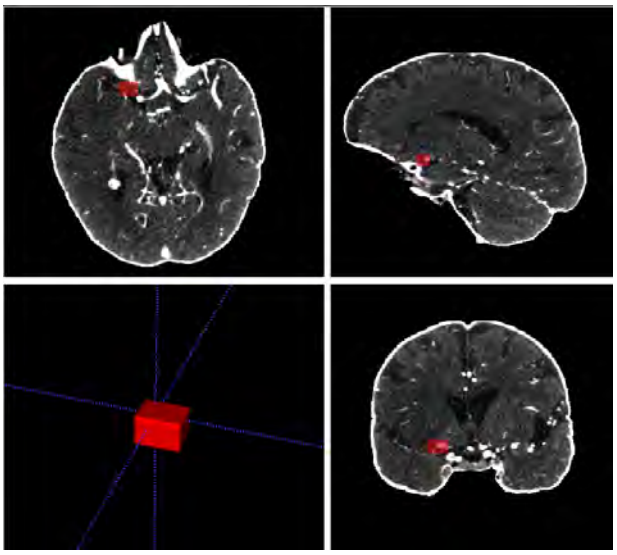
3D object detection involves estimating an object's location, class, orientation, and depth. 2D to 3D object detection methods rely on 2D image data and estimate depth and orientation for 3D bounding boxes. 3D object detection models with volumetric input directly process 3D representations, providing more detailed scene information. While both approaches have advantages, the latter generally achieves higher accuracy but requires more computationally intensive processing. For instance, Mono3D (He and Soatto, 2019), which converts 2D detection results to 3D bounding boxes, achieves a mean average precision (mAP) of 70.1% on the KITTI dataset. In comparison, VoxelNet (Zhou and Tuzel, 2017), a 3D object detection model with volumetric input, achieves an mAP of 73.4% on the same dataset.

3D detection has been applied in medical image processing for automatic early diagnosis and screening. For example, (Zhu et al., 2018) and (Xie et al., 2019) utilize 3D CT and MRI imagery. Numerous studies have explored 3D detection in medical imagery. Hu et al. (2018) reviewed recent works on medical image based cancer detection and diagnosis, most of which have employed 3D CNN schemes for detection. Monkam et al. (2019) reviewed the advancement of detection and classification of pulmonary nodules using 3D CNN in CT imagery. 3D CNN frameworks such as 3D U-Net (Tang et al., 2018) and 3D DenseNet, 3D Faster R-CNN have been employed for 3D nodule detection. These frameworks have improved accuracy by employing ensembles of multiple CNN models and fine-tuning hyperparameters. An example of such a network is DeepLung by Zhu et al. (2018), which achieved an impressive 90.4% accuracy in pulmonary nodule detection, significantly outperforming 2D object detection methods. Also, there is the 3D detection network 3D CNN, motivated by the work of Feature Pyramid Networks. it is 3D nodule detection and has won the 1st place team of Kaggle Data Science Bowl competition (Liao and Song., 2019) with a 86% sensitivity and false positive rate of 8 false positives per scan. Despite these advancements, training data limitations remain a challenge for 3D CNN-based detection. Collecting, storing, and annotating 3D data is more complex compared to 2D images. Therefore, existing datasets are relatively small, hindering the development of more robust models.

Our analysis of existing literature underscores the advantages of employing 3D object detection models, indicating significant enhancements in outcomes. Specifically, pioneering models like nnDetection and 3D CNN frameworks have exhibited promising efficacy in streamlining the automated detection process for LVOs. Consequently, our research endeavors will be focused on delving into the practical implementation of these advanced models to proficiently identify and characterize LVOs within CTA images.



(a) Change in contrast highlights the region of a large vessel occlusion (LVO) in a CTA image, illustrating the challenges in accurately identifying the occlusion boundaries



(b) 3D bounding box applied to a CTA brain image, representing the area of the large vessel occlusion (LVO) with a 2-pixel margin, providing a practical solution to segmentation difficulties

Figure 3: The challenges associated with obtaining precise 3D segmentation annotations of large vessel occlusions in a CTA dataset

### *2.3. Previous work*

In prior research conducted within the VICOROB Research Group, Paola Martinez Arias (2023) explored the detection and classification of Large Vessel Occlusions (LVOs) using the same dataset employed in our study. Martinez's investigation, completed in 2023, focused on two primary tasks: binary classification of CTA images to determine the presence or absence of LVOs, and precise localization of occlusions through 3D bounding boxes.

For the classification task, Martinez conducted a comparative analysis of utilizing brain symmetry information against models trained without such data. Her findings revealed that incorporating symmetry information significantly enhanced model performance, with top-performing experiments achieving a remarkable 77% accuracy in inferring LVO presence on unseen hospital datasets. In parallel, for the occlusion detection task, Martinez leveraged the nnDetection framework, training the model on three folds of hospital data. The outcomes were promising, demonstrating robust detection capabilities in both anterior and posterior circulation occlusions. Notably, the detector achieved a sensitivity of 97% on test cases, coupled with an impressively low false positive per image (FPpI) rate of approximately 0.15. These findings underscore the efficacy of advanced DL models in accurately detecting and characterizing LVOs, providing a solid foundation for our ongoing research endeavors.

## 3. Material and methods

### *3.1. Data*

This study utilized a dataset of Computed Tomography Angiography (CTA) scans acquired from Hospital Dr. Josep Trueta in Girona, Spain. The dataset initially comprised 321 cases involving patients diagnosed with an LVO. The distribution of occlusions across various locations within the brain vasculature provided valuable insights into the prevalence of LVOs in different segments. Here's a breakdown of the distribution:

- M1 segment: 46%
- M2 segment: 18%
- ICA (Internal Carotid Artery): 13%
- Tandem (combined M1 and M2 segment occlusion): 7%
- Basilar artery: 6%
- PCA (Posterior Cerebral Artery): 5%
- Other locations (M3, A2, VA, and extracranial): 5%

After excluding cases with 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. The resolution of the CTA scans was 512x512x378, with variable voxel spacing and a slice thickness of 0.9. All of the examinations were performed with a Philips Healthcare Ingenuity CT scanner. Out of the 310 valid cases, only 124 were ultimately

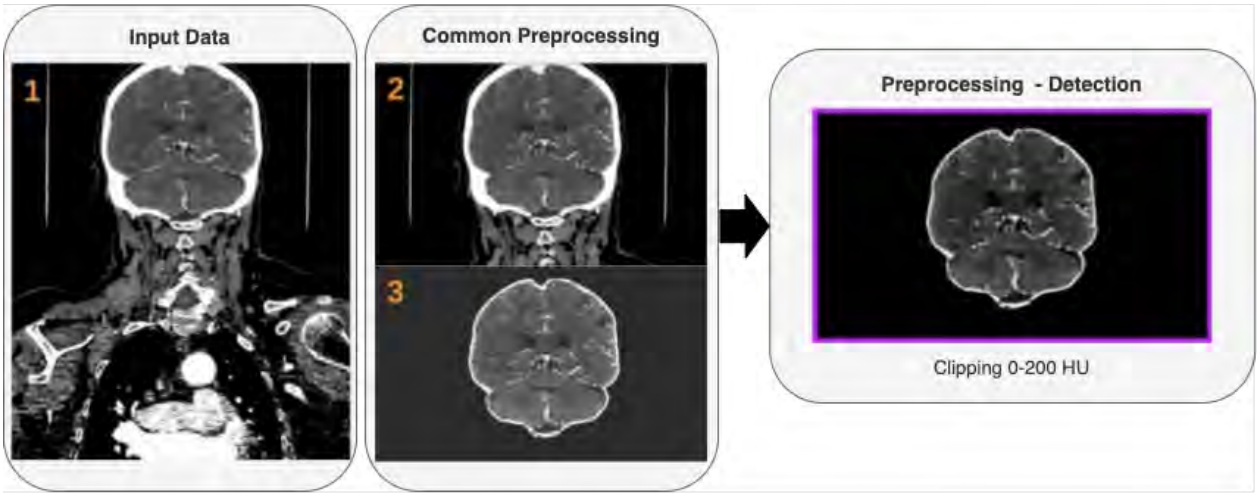



Figure 4: Preprocessing Pipeline. 1. Image in nifti format. 2. Results after using *robustfov*. 3. Results after using bet. In the final step, we perform clipping of intensities.

used for the LVO detection task due to the availability of ground truth thrombus segmentation obtained by manual annotations using ITK-snap. These annotations were carried out by expert technicians, with the guidance and validation of all cases from an expert neurologist at the collaborating Hospital. The annotations were focused on the main slices where the thrombus size and shape were most discernible, rather than on all slices of the CTA scans.

Due to the significant challenges associated with obtaining precise 3D segmentation annotations across all three views of our CTA dataset (as illustrated in Fig. 3), we opted for 3D bounding boxes to represent the LVO regions. These challenges include the complex and variable anatomy of blood vessels, the presence of noise and artifacts in the imaging data, and the time-consuming nature of manual segmentation, which often leads to inconsistencies across different views and slices. These bounding boxes encompass the clot in the CTA image, providing a simplified yet informative approximation of the area where the occlusion is most likely to occur. we created the bounding boxes around the manually annotated thrombus segmentation, with a margin of 2 pixels for all cube coordinates. Each case in our dataset features exactly one occlusion.

This detailed explanation of the data section provides a clearer understanding of the dataset's characteristics, the selection process, and the rationale behind ground truth annotation choices.

### *3.2. Data pre-processing*

To preprocess the hospital dataset, we followed several steps, as shown in Fig. 4.

Initially, we converted the data from DICOM to NIFTI format. Subsequently, we used FSL (Jenkinson et al., 2012) for further preprocessing. The first step involved applying *robustfov* to focus on the skull in the images, given that the original images included not only the head but also the patient's upper body. Following this, we employed BET for skull stripping. Obtaining, in the end, just the brain in the CTA images. In the final step, we clipped the intensity values of the brain between 0-200 HU, based on the suggestions from hospital doctors. The images were neither registered nor resized, maintaining their original dimensions for input into our models.

### *3.3. Methodology*

In our study, we analyze two distinct approaches for detecting LVOs: nnDetection and a 3D CNN.

nnDetection is a cutting-edge 3D self-configuring medical object detection model developed by Baumgartner (2021). It adapts itself without any manual intervention to arbitrary medical detection problems while achieving results en par with or superior to the state-of-the-art. The model's effectiveness has been demonstrated on two public benchmarks, such as ADAM and LUNA16, and it has been evaluated on ten additional public datasets to assess its comprehensive performance in medical object detection.

3D CNN for detection is a network originally designed for detecting lung nodules. This 3D nodule network is inspired by Feature Pyramid Networks and was part of the winning entry in the Kaggle Data Science Bowl competition (Liao and Song., 2019).

The following sections provide a detailed description of these two approaches, including their architecture

and training plans, followed by an outline of the evaluation metrics used.

### *3.3.1. nnDetection for 3D object detection*

The first approach we employed was the nnDetection framework, a state-of-the-art 3D self-configuring medical object detection model. The framework facilitates adaptation to new datasets and utilizes a 5-fold cross-validation system for data splitting and training plan creation.

The input to nnDetection consists of a stack of 3D images from a CTA scan and the corresponding 3D bounding boxes. The model output consists of a dictionary for each test case with the following information: predicted bounding boxes, prediction scores, predicted labels, original size of the raw data, origin of the image as read by the Insight Toolkit (ITK), ITK spacing and ITK direction.

*Architecture.*

nnDetection selects the best architecture depending on the dataset by adhering to a set of interdependent principles: (1) data fingerprint, which covers the relevant properties of the training data; (2) rule-based parameter, which employs a set of heuristics based on the fingerprint; (3) fixed parameters, which do not rely on the data; and (4) empirical parameter optimization, which is the set of parameters optimized during the training.

The detection algorithm is based on Retina-UNet (Jaeger et al., 2020), which combines the RetinaNet one-stage detector with the U-Net architecture commonly used for semantic segmentation. This combination enhances object detection with semantic segmentation capabilities without introducing additional complexity. Fig. 5 demonstrates the schematic representation of the baseline topology used in the present study.

*Training plan.*

The nnDetection framework automatically generates a training plan tailored to the dataset. The plan consists of several steps, including cropping, preprocessing of the input images, and details about the architecture and the input sizes of the images. The model training involves a sum of cross-entropy (classification) and generalized intersection-over union (regression) loss, carried out in a five-fold cross-validation setup to differentiate between background and labeled LVOs.

Training was performed for 60 epochs using stochastic gradient descent with Nesterov momentum of 0.9. Initially, the learning rate was linearly increased from 1e−6 to 1e−2 over the first 4000 iterations. A polynomial (poly) learning rate schedule was applied until epoch 50, gradually decreasing the learning rate to ensure stable convergence. In the final 10 epochs, a cyclic learning rate fluctuating between 1e−3 and 1e−6 was used, which helps the model escape local minima and potentially find a more optimal set of parameters.

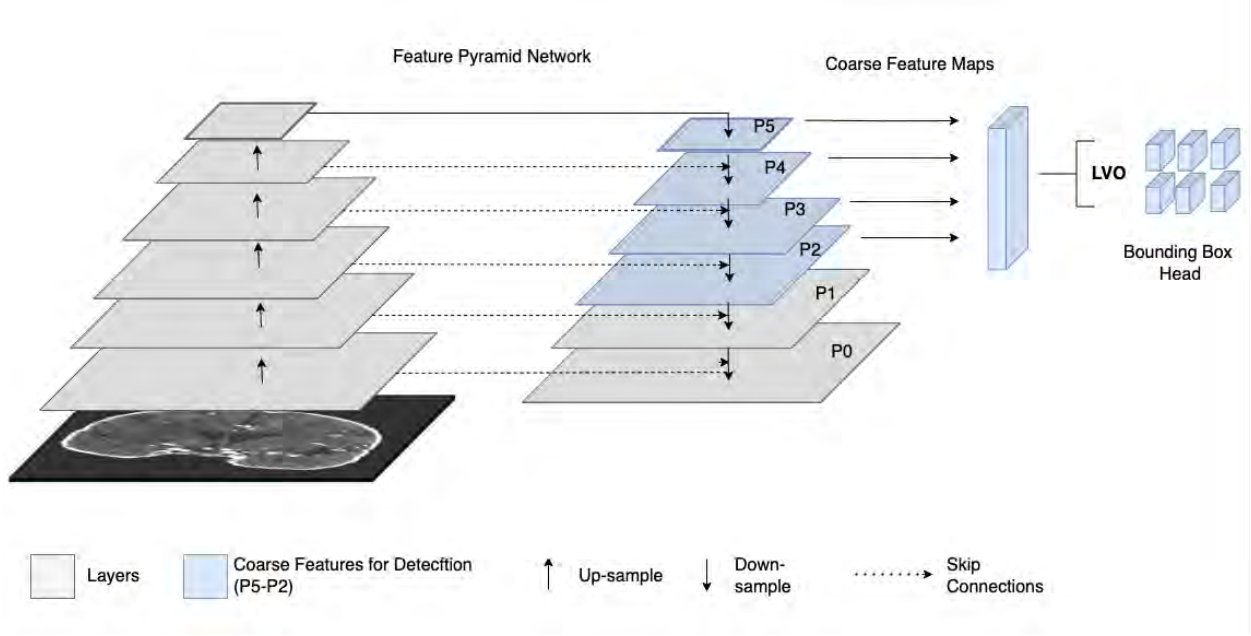



Figure 5: The backbone of nnDetection is a pyramid-like network with bottom-up (left) and top-down (right) pathways with interconnected layers. The upper layers have a lower spatial resolution yet have representative features for the task at hand. Classification and regression tasks (e.g., bounding box determination) are performed on averaged feature maps. On the bottom-up path, the spatial resolution of the feature maps decreases while getting richer and denser information, while the top-down path recovers the spatial dimension. The skip connections facilitate information exchange between pathways. The blue-colored features maps (P5–P2) are utilized for LVO object detection.

This phase also incorporates Stochastic Weight Averaging (SWA), which averages the weights from different points in the learning rate cycle, thus enhancing the model's generalization ability and improving its overall performance. Training was performed on patches to overcome the memory limitations caused by the 3D model configuration; the patch size used is [160, 160, 128], with a target spacing of [0.45001221 0.5859375 0.5859375 ] and a batch size of 4, sampled from the CTA scans while ensuring an equal number of foreground and background patches per batch.

*Data Augmentation.*

The same data augmentation strategy of nnU-Net (Isensee et al., 2020), which includes Gaussian noise, Gaussian blur, brightness, contrast, simulation of low resolution, gamma augmentation, elastic deformation, scaling, flipping, mirroring, and rotation, was implemented during the training process.

### *3.3.2. 3D CNN for 3D object detection*

The second approach we tested for detecting LVOs in CTA scans is a 3D CNN, adapted from a model originally designed for detecting lung nodules (Liao and Song., 2019). This network leverages a 3D U-Net backbone and a Region Proposal Network (RPN) output layer to effectively identify and localize LVOs.

The input to the network consists of a stack of 3D images from a CTA scan and the corresponding information of 3D bounding boxes (Gx,Gy,Gz,Gr), where the first three elements denote the coordinates of the center point of the box and the last element denotes the depth.

*Architecture.*

The network architecture consists of a 3D U-Net backbone (Çiçek et al., 2016) and an RPN output layer, and its structure is shown in Fig. 6. The U-Net backbone enables the network to capture multiscale information, which is essential for detecting LVOs of various sizes. The architecture employs several residual blocks to enhance feature extraction, thereby improving the overall performance. The RPN output layer is used to directly generate object proposals, streamlining the detection process.

*Training.*

To compare with nnDetection, we conducted 5-fold cross-validation experiments with the same split

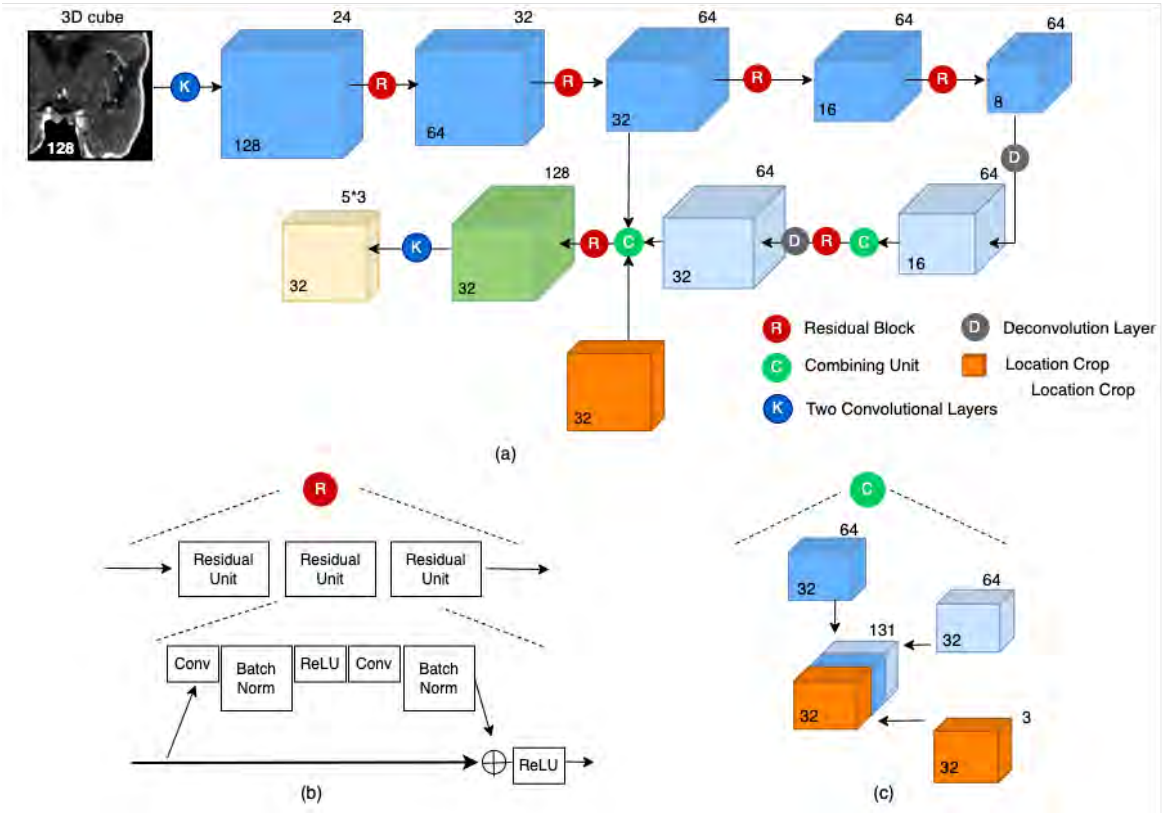



Figure 6: Detection net. (a) Overall network structure. Each cube in the figure stands for a 4-D tensor. Only two dimensions are indicated in the figure. The number inside the cube stands for the spatial size (Height = Width = Length). The number outside the cube stands for the number of channels. (b) Structure of a residual block. (c) Structure of the left combining unit in (a). The structure of the right combining unit is similar but without the location crop. Best viewed in color.

as nnDetection. The training process involves extracting small 3D patches from the CTA scans and feeding them individually into the network. These patches have dimensions of 128×128×128×1 (Height×Length×Width×Channel). The selection process involves randomly sampling two types of patches. Approximately 70% of the input contain at least one LVO, while the remaining 30% are cropped randomly from the scans and may not include any LVOs. This strategy ensures that the training data includes enough negative samples, which is crucial for robust learning.

Patches are padded with a constant value and augmented through random flipping and resizing. This augmentation helps the model generalize better by exposing it to a variety of transformations.

*Location Information and Loss Function.*

To aid in the identification and localization of LVOs, location information is introduced into the network. A smoothed L1-norm function is used as the loss metric for bounding box regression, which provides robust optimization and improves accuracy. To address class imbalance during training, positive sample balancing is employed, along with hard negative mining techniques.

*Inference.*

After training, the test cases in each fold are split into several parts (208×208×208×1 per part) to overcome GPU memory constraints during testing. The results from these parts are then combined to obtain the final detection outputs. A non-maximum suppression operation is performed to eliminate overlapping proposals and generate a more refined set of potential LVO locations.

#### *3.3.3. Evaluation metrics.*

For the detection problem in our study, we considered Intersection over Union (IoU), sensitivity, and false positive per image rate (FPpI), all based on True Positives (TP), False Positives (FP), and False Negatives (FN). Intersection over Union (IoU) measures the overlap between the predicted and 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 prediction correct if it achieves an IoU of at least 0.1. This threshold respects the clinical need for coarse localization and leverages the non-overlapping nature of objects in 3D, as noted by Jaeger et al. (2020).

To calculate the number of TP, FP, and FN, we considered different confidence thresholds and a minimum IoU of 0.1 with the ground truth for a prediction to be classified as TP. Since all cases in our dataset contain exactly one occlusion, we simplified the TP identification by selecting the bounding box with the highest confidence score in cases where multiple TPs are found. For FP, we counted them as 1 if the bounding boxes showed a large intersection between them. Additionally, we considered discarding FP if they did not overlap with the 3D vasculature representing the Circle of Willis (Lal-Trehan Estrada et al., 2024). An FN is any case where the predicted bounding box had an IoU less than the minimum IoU and a confidence score below the threshold used for TPs.

In summary, our evaluation metrics are designed to balance the need for precise localization of LVOs with the practical considerations of working with 3D medical imaging data. By setting appropriate thresholds and refining our FP criteria, we aim to ensure that our detection model meets the clinical requirements for identifying LVOs in CTA scans.

## 4. Results

In this section, we present the performance evaluation of the two approaches: nnDetection and 3D CNN.

### *4.1. nnDetection*

In this subsection, we present the results of nnDetection over 5 folds. As mentioned in Section 3, we performed the experiments using a 5-fold cross-validation on 124 valid cases, with 25 cases allocated for testing in each fold. The overall results of the nnDetection model on the testing cases are illustrated in Fig. 7, providing insights into the model's performance across different thresholds and after false positive reduction.

Fig. 7a displays two curves representing the True Positive Rate (TPR) and False Positives per Image (FPpI) rate at various confidence thresholds. The blue curve illustrates the global TPR/FPpI before reducing false positives, while the red counterpart demonstrates the TPR/FPpI post-reduction. Impressively, both curves exhibit a commendable sensitivity of approximately 90% sensitivity. However, while the global curve indicates over 2 FP per image, the red curve reduces this to below 2 FP per image.

The second figure provides a more granular view of TPR and FPpI rates across different confidence score thresholds. Here, the blue plot represents the TPR before and after FP reduction, shedding light on the model's sensitivity across varied thresholds. Meanwhile, the green plot signifies the FPpI rate in the global context, and the red plot depicts the FPpI rate post-false positive reduction. A subtle variance is discernible between the two, notably prominent at lower confidence scores (e.g., 0.1, 0.2).

*Performance Metrics*

Further dissecting the model's performance, we meticulously analyzed the True Positive Rate (TPR) and the rate of False Positives per Image (FPpI) metrics at distinct confidence thresholds, as shown in Fig. 7b. Lower thresholds (e.g., 0.1, 0.2) resulted in higher true positive rates but also increased false positives. Conversely, higher thresholds (e.g., 0.8, 0.9) reduced false positives but also decreased true positives. The optimal threshold, striking a balance between these metrics, was identified at 0.6, yielding approximately 0.14 FPpI and 71% sensitivity.

*Cross-Validation Results*

Table 1 summarizes the results for each fold employing the optimal Confidence Threshold of 0.6. Sensitivity spanned from 0.56 to 0.84, while FPpI rate exhibited variance from 0.04 to 0.28. The mean sensitivity across folds stood at 0.71, with a standard deviation of 0.10, while the mean FPpI registered at 0.18, with a standard deviation of 0.11.

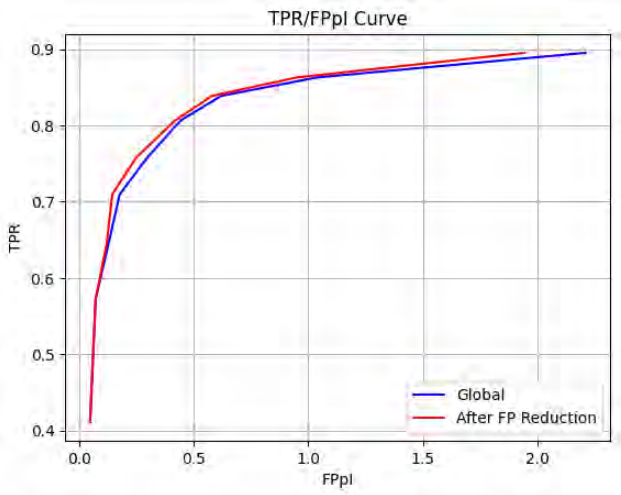
| Fold | TP, FP, FN counts | Sensitivity | FPpI |
|---|---|---|---|
| Fold 0 | TP: 14, FP: 7, FN: 11 | 0.56 | 0.28 |
| Fold 1 | TP: 18, FP: 6, FN: 7 | 0.72 | 0.24 |
| Fold 2 | TP: 19, FP: 1, FN: 6 | 0.76 | 0.04 |
| Fold 3 | TP: 21, FP: 7, FN: 4 | 0.84 | 0.28 |
| Fold 4 | TP: 16, FP: 1, FN: 8 | 0.67 | 0.04 |
| Mean | | 0.71 | 0.18 |
| Std | | 0.10 | 0.11 |

Table 1: Evaluation Results with Confidence Threshold of 0.6

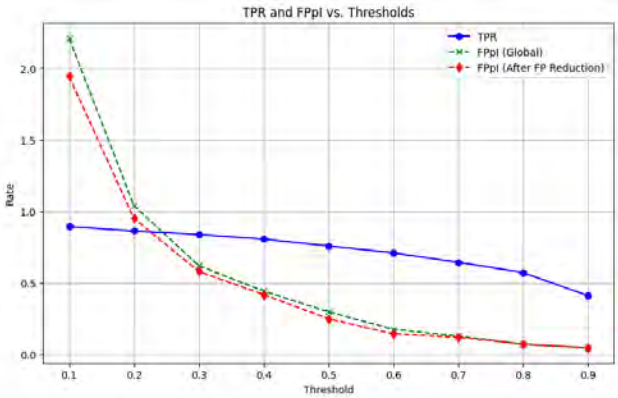
*False Positive Reduction*

In our endeavor to refine the results and by knowing that LVOs exist in the Circle of Willis, we further analyzed the false positives by assessing their intersections with the 3D vasculature of the Circle of Willis. False positives devoid of such intersections were discarded, resulting in a reduction in false positives, particularly noticeable at lower thresholds (e.g., from 274 to 241 at a confidence threshold of 0.1), as shown in the red plot in Fig. 7b. This in-depth analysis underscores the model's proficiency in making predictions within the region of interest, thereby enhancing its clinical relevance and accuracy.

Through this comprehensive examination, we glean valuable insights into the nnDetection model's performance, thereby facilitating the optimization of its effectiveness in detecting LVOs.




(a) TPR/FPpI curve across different confidence thresholds



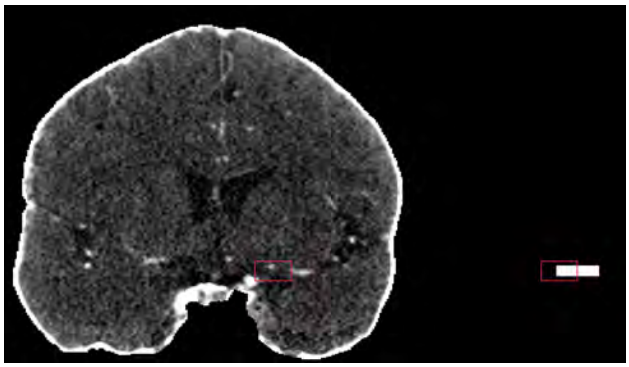

(b) TPR and FPpI rate at different confidence thresholds

Figure 7: Results of nnDetection, including before and after FP reduction

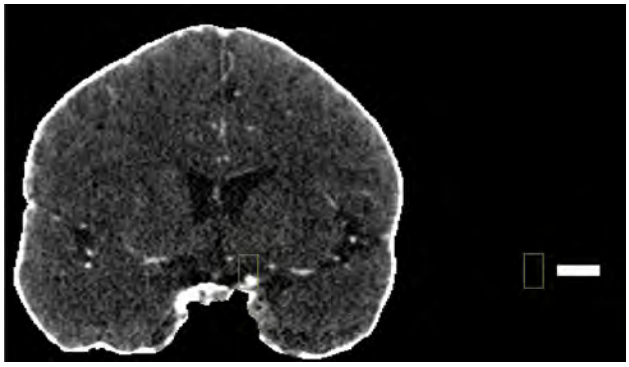
*Illustrative Examples of Detection Performance*

The nnDetection model's performance is visually assessed through qualitative examples depicted in Fig. 8. Here, we provide a nuanced examination of the model's ability to accurately identify occlusions in various scenarios.

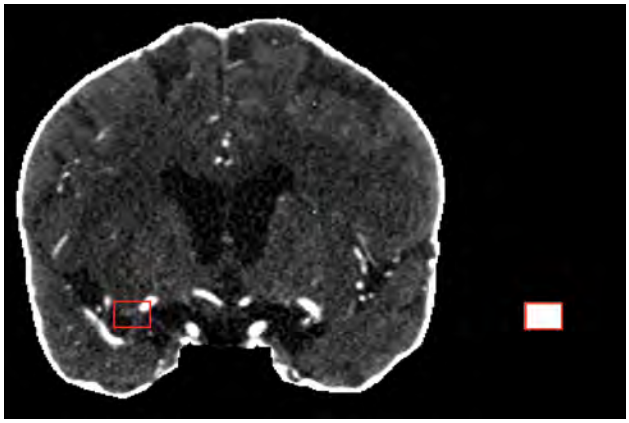
In Fig. 8a, we observe an example of true positive detection, where the red bounding box overlaps with the occlusion denoted by the white bow, representing the ground truth. Conversely, Fig. 8b illustrates a false positive detection, where the green bounding box deviates from the ground truth. Upon visual inspection, it becomes apparent that subtle differences in contrast along the arteries can be misconstrued as occlusions, leading to false positive detections. This underscores the importance of careful scrutiny and validation of model outputs to mitigate potential mislocalization and ensure the reliability of the detection process.



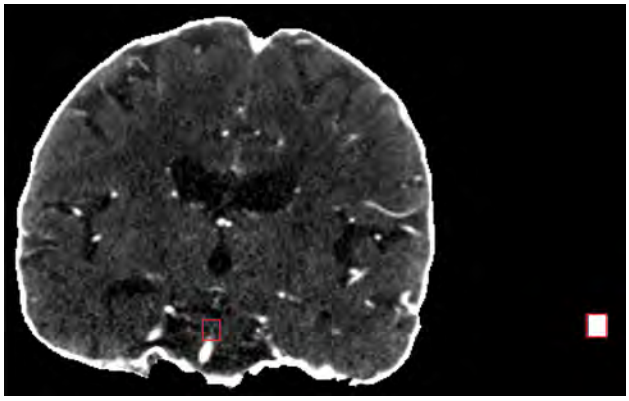
(a) True Positive detection.



(b) False Positive detection



(c) True Positive detection on an M2 case.



(d) Inference over a posterior circulation image

Figure 8: Bounding boxes results of detection with a prediction score of 0.6. The white box located in the right represents the ground truth of the occlusion

Moving forward, Fig. 8c showcases true positive detection in an M2 case, where the model adeptly identifies the occlusion with precision, demonstrating its efficacy in discerning subtle nuances within the images. Lastly, Fig. 8d highlights true positive detection in a posterior circulation image, wherein the model accurately detects the occlusion despite potential challenges posed by limited training data. This underscores the robustness of the nnDetection model in identifying occlusions across diverse anatomical configurations.

From the images, we can conclude that the occlusion is precisely detected by the generated bounding box. However, as illustrated in Fig. 8b, 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, which can be confused with an LVO.

This qualitative assessment provides useful insights into the nnDetection model's performance, revealing its strengths and areas for improvement in effectively detecting occlusions in CTA scans.

### *4.2. 3D CNN*

The initial evaluation of the 3D CNN approach yielded unsatisfactory results. The model identified a large number of potential LVOs, leading to a high false-positive (FP) to true-positive (TP) ratio. This limited the model's practical utility. Figure 9 illustrates an example where numerous positive detections overwhelm the image, making it difficult to distinguish true positives from false positives.

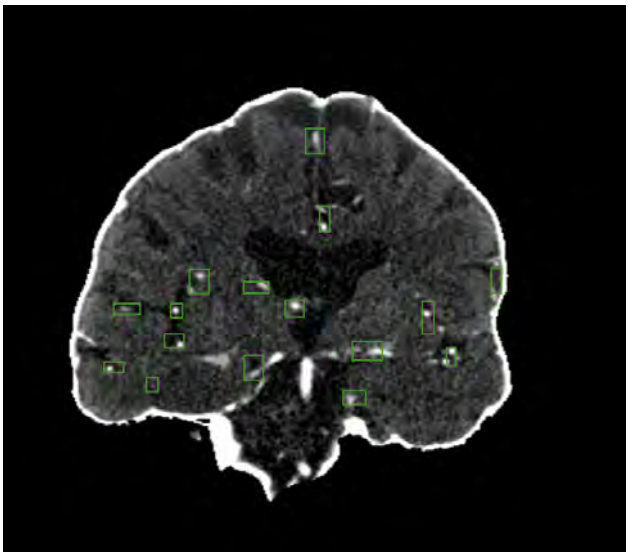


Figure 9: Example of excessive false positive detections by the 3D CNN model in a single case.

*Improvement Strategies*

To address the high false-positive rate, we propose several strategies:

- **Dual 3D CNN Classifiers:** Utilizing two classifiers to distinguish false positives more accurately.
- **Loss Function:** Replacing the cross-entropy loss function with a focal loss function to handle class imbalance effectively.
- **Data Augmentation:** Increasing the number of positive samples through oversampling techniques such as sliding window cropping, flipping (x-axis, y-axis, z-axis), rotation (90°, 180°, 270°), and multi-scale transformation.

These improvements are expected to enhance the utility of the 3D CNN model, making it a more competitive approach for LVO detection.

### *4.3. Inference on IACTA-EST Challenge dataset*

To assess the generalizability of the nnDetection model, we conducted inference on 20 cases from the first task of the IACTA-EST Challenge dataset. This dataset is entirely independent of the Hospital Dr. Josep Trueta dataset used for model training.

The Image Analysis for CTA Endovascular Stroke Therapy (IACTA-EST) Challenge provides a valuable resource by offering a curated image dataset from multiple clinical sites. This approach aims to bridge the gap between current research and commercially available solutions by incorporating diverse data sources.

The 20 cases used for inference comprised 10 LVO-positive and 10 LVO-negative examples. Notably, this data originated from different domains and utilized scanners distinct from those employed in the training dataset. Furthermore, no preprocessing steps were applied to these cases; inference was performed directly.

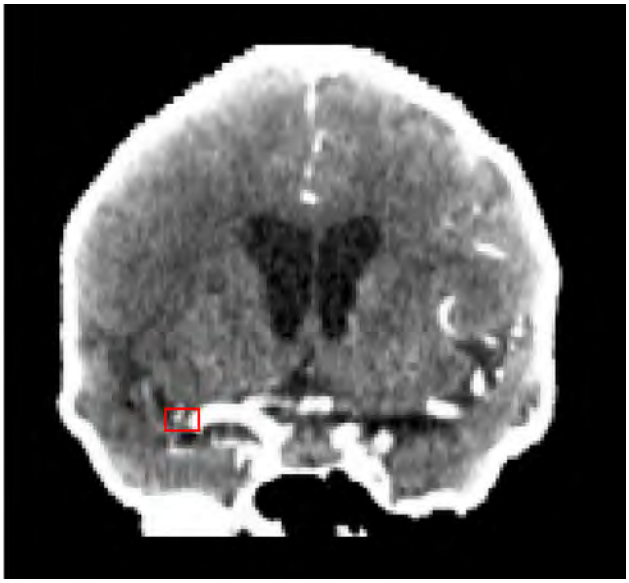
*Overall Performance and Generalizability Insights*

The inference results revealed slightly more positive class predictions across various confidence thresholds. This suggests that the model might have a tendency to favor identifying occlusions even when they may not be present. Notably, at a threshold of 0.3, no negative predictions were observed, while two positive cases yielded higher positive scores. These findings highlight the potential for improvement in model generalizability.

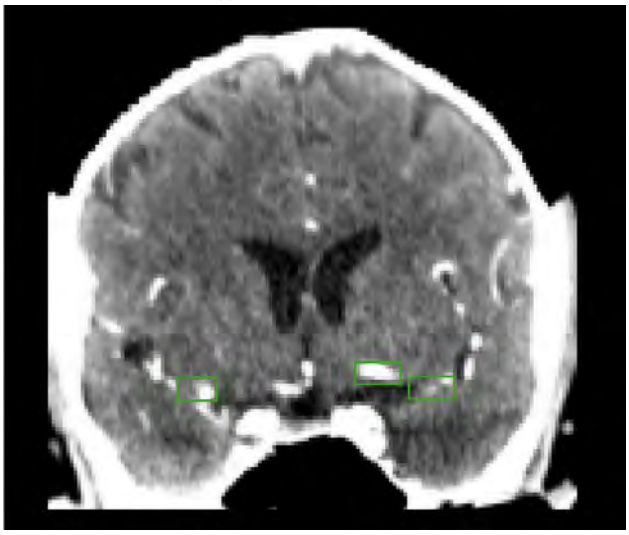
*Qualitative Examples*

Fig. 10a showcases a successful prediction of an LVO from a positive case, where the model adeptly identifies the area of occlusion. The predicted confidence score exceeded 0.9, indicating a strong likelihood of occlusion. This demonstrates the model's ability to accurately identify LVOs in unseen data under certain circumstances.

Conversely, Fig. 10b depicts a false positive prediction for a negative case, where there are multiple false positives, with the green bounding boxes highlighting certain areas where there are occlusions but no LVO in this example. The model incorrectly predicted the presence of an LVO by assigning a confidence score of 0.2. This example demonstrates the need for additional refinement to improve the model's ability to distinguish between true and false positives, particularly when encountering data from diverse sources.



(a) Positive case detection.



(b) Negative case detection

Figure 10: Inference results of two cases from the IACTA-EST dataset. The white box located in the right represents the ground truth of the occlusion

The inference results on the IACTA-EST dataset highlight the need for improved generalizability for real-world use. To address this, we plan to explore fine-tuning the nnDetection model on a combined dataset

incorporating both the Hospital Trueta data and a subset of IACTA-EST data. This can improve the model's ability to adapt to data variations. Alternatively, we can leverage transfer learning, using knowledge from the pre-trained nnDetection model to train a new model specifically for the IACTA-EST Challenge data. These refinements aim to equip the nnDetection model with a more robust ability to generalize to diverse data sources, ultimately enhancing its clinical applicability.

## 5. Discussion

### *5.1. nnDetection*

This study investigated the effectiveness of the nnDetection framework for precise 3D localization of occlusions in CT angiograms (CTAs). The results demonstrated that the trained nnDetection model could accurately detect true positives (TPs) with a very low false positive per image (FPpI) rate. As illustrated in Fig. 7b, the model performs best at a confidence threshold of around 0.6, balancing TP detection while minimizing false positives (FPs).

#### *Analysis of Occlusion Localization*

The analysis of occlusion localization in the test cases yielded several key insights. The training dataset was primarily composed of M1 cases, with fewer instances of M2, ICA, Tandem, Basilar, and PCA occlusions. The 5-fold cross-validated models successfully detected occlusions in M2, ICA, and Tandem cases. There were some false positives noted in M1 cases, but their overall occurrence was relatively low. Importantly, all occlusions in the test cases belonged to the Anterior Circulation system.

Despite a limited number of Posterior Circulation cases in the training set, the model demonstrated its capability to detect Basilar occlusions during inference, as shown in Fig. 8d. This finding is particularly encouraging, suggesting that the model can generalize beyond the predominantly anterior circulation training data. These promising results advocate for further validation studies involving a larger and more diverse dataset and with different types of occlusions.

#### *Computational Considerations*

A significant drawback of the nnDetection model is the substantial computational time required for training and inference. For the dataset comprising approximately 100 CTAs, each fold necessitated around 6 days of training. Additionally, the prediction and inference processes for each image required about 5 minutes, excluding preprocessing and cropping performed automatically by the framework. This extensive computational demand necessitates the development of optimized strategies to enhance efficiency and reduce training time.

### *5.2. 3D CNN Model Challenges*

The 3D CNN detection model encountered significant challenges, particularly regarding computational resources. Training and testing necessitated all three available GPUs, each with 12 GB of memory, and a batch size of 8. The initial results were unsatisfactory, with a high false positive (FP) to true positive (TP) ratio of 1766166:582357 before discarding overlapping predictions.

To address these issues, several strategies are planned:

- **Dual 3D CNN Classifiers:** Employing two classifiers to more accurately distinguish false positives.
- **Loss Function Adjustment:** Replacing the cross-entropy loss function with a focal loss function to better handle class imbalance.
- **Data Augmentation:** Increasing the number of positive samples through oversampling techniques such as sliding window cropping, flipping (along the x-axis, y-axis, and z-axis), rotation (90°, 180°, 270°), and multi-scale transformation.

These improvements are anticipated to enhance the detection performance of the 3D CNN model significantly.

### *5.3. Future Directions*

While this study demonstrates the promise of the nnDetection framework for accurate occlusion detection in CT angiography (CTA) images, significant advancements are necessary to address its computational demands and the initial challenges encountered with the 3D convolutional neural network (3D CNN) model.

Future efforts will focus on optimizing both models to achieve superior performance while minimizing resource consumption. This will involve exploring advanced training techniques, efficiently utilizing computational resources, and significantly expanding the dataset to encompass a broader spectrum of occlusion types and anatomical variations. This expansion includes investigating the model's ability to predict the location of Large Vessel Occlusions (LVOs) beyond the currently identified common cases like M1, M2, ICA, and even including occlusions in the posterior circulation. Additionally, we will investigate incorporating prior knowledge, such as segmentation masks of the Circle of Willis (CoW), to guide the detection process and potentially improve accuracy. Finally, the model's generalizability will be rigorously evaluated using data from different hospitals and scanner types, and potentially further enhanced by incorporating data annotated

by multiple neurologists with expertise in CTA image interpretation.

This multi-expert approach can help account for inter-rater variability in occlusion detection and potentially lead to a more robust and reliable model. By addressing these future directions, the nnDetection framework has the potential to evolve into a robust, versatile, and clinically valuable tool for accurate LVO detection in CTA images.

## 6. Conclusions

This study investigated the efficacy of two deep learning approaches, nnDetection and 3D CNN, for detecting Large Vessel Occlusions (LVOs) in stroke patients using Computed Tomography Angiography (CTA) images. The study provides a comprehensive analysis of each approach's performance, highlighting their strengths, limitations, and potential areas for future development.

The nnDetection framework demonstrates promising results in accurately localizing occlusions, particularly within the Anterior Circulation system. This suggests its potential for clinical applications. However, a major drawback is the significant computational cost associated with both training and using the model (inference time). Despite these challenges, the model shows potential for generalizing to different types of occlusions beyond those included in the training dataset.

While the 3D CNN model offers a different approach to LVO detection, it encountered significant challenges, primarily related to high computational demands and a large number of false positive results. The study proposes several strategies to address these issues, including utilizing dual classifiers, adjusting loss functions, and enriching the training data with data augmentation techniques.

Encouragingly, the nnDetection framework achieved very positive results. We are particularly enthusiastic about the feedback from a collaborating neurologist who assessed the visual analysis of the nnDetection results. They reported that the model's detection capabilities surpassed those of currently used commercial software. This suggests that the nnDetection framework has the potential to become a valuable clinical tool for stroke diagnosis and treatment planning.

Future research directions include exploring advanced training techniques, optimizing computational strategies to improve efficiency, and significantly expanding the dataset to encompass a broader spectrum of occlusion types and anatomical variations. This expansion could involve incorporating data from various sources, such as different hospitals and scanner types, to enhance the model's generalizability. Ultimately, these efforts aim to develop more efficient and accurate deep learning tools for early stroke detection and treatment. By facilitating timely intervention, these tools have the potential to significantly improve patient outcomes and reduce the overall healthcare burden associated with stroke.

## Acknowledgments

I would like to express my gratitude towards my supervisors, Dr. Xavier Llado and Dr. Arnau Oliver, for their invaluable guidance, encouragement, and support throughout this project. Also, I would like to thank the VICOROB research group for their help, especially Valeriia Abramova and Uma Maria Lal-Trehan Estrada for their contribution to this work. I am also grateful to Dr. Mikel Terceno for providing the dataset essential to this research and for his patience and willingness to teach and help with the annotations. Additionally, I would like to thank the MaIA consortium and the friends I have made during this journey.

Lastly, I am forever thankful to my family, friends, and loved ones for their unwavering support and encouragement over these two years. This achievement would not have been possible without you.

## References

Bagcilar, O., A.D.A.C.e.a., 2023. Automated lvo detection and collateral scoring on cta using a 3d self-configuring object detection network: a multi-center study. Sci Rep 13, 8834. doi:https://doi.org/10.1038/s41598-023-33723-w.

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, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 1873–1877. doi:10.1109/ISBI.2019.8759475.

Baumgartner, Paul F. Jäger, F.I..K.H.M.H., 2021. nndetection: a self-configuring method for medical object detection, in: Miccai 2021: 24th international conference, strasbourg, france. Springer 24, 530–539.

Bruggeman, 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. doi:10.1161/SVIN.121.000158.

Brugnara, Baumgartner M, S.E.e.a., 2023. Deep-learning based detection of vessel occlusions on ct-angiography in patients with suspected acute ischemic stroke. Nat Commun 14(1), 4938. doi:doi:10.1038/s41467-023-40564-8.

Chavva, Crawford AL, M.M.e.a., 2022. Deep learning applications for acute stroke management. Ann Neurol 92(4), 574–587. doi:doi:10.1002/ana.26435.

Czap, Bahr-Hosseini M, S.N.e.a., 2022. Machine learning automated detection of large vessel occlusion from mobile stroke unit computed tomography angiography. Stroke 53(5), 1651–1656. doi:10.1161/STROKEAHA.121.036091.

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:https://doi.org/10.1016/j.nicl.2023.103362.

He, T., Soatto, S., 2019. Mono3d++: Monocular 3d vehicle detection with two-scale 3d hypotheses and task priors. ArXiv abs/1901.03446.

Hu, Z., Tang, J., Wang, Z., Zhang, K., Zhang, L., Sun, Q., 2018. Deep learning for image-based cancer detection and diagnosis-a survey. Pattern Recognit. 83, 134–149.

Isensee, F., P.F.J., Kohl, S.A.A., Petersen, J., Maier-Hein, K., 2020. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18, 203 – 211. URL: https://api.semanticscholar.org/CorpusID:227947847.

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.J., Woolrich, M.W., Smith, S.M., 2012. Fsl. NeuroImage 62, 782–790. URL: https://api.semanticscholar.org/CorpusID:208816469.

JoeNiekroFoundation, 2017. Brain basics URL: joeniekrofoundation.com/understanding/brain-basics/.

Lal-Trehan Estrada, U., Oliver, A., Sheth, S.A., Lladó, X., Giancardo, L., 2024. Strategies to combine 3d vasculature and brain cta with deep neural networks: Application to lvo. iScience 27, 108881. doi:https://doi.org/10.1016/j.isci.2024.108881.

Liao, M. Liang, Z.L.X.H., Song., S., 2019. Evaluate the malignancy of pulmonary nodules using the 3-d deep leaky noisy-or network. IEEE Transactions on Neural Networks and Learning Systems 30, 3484–3495. doi:10.1109/TNNLS.2019.2892409.

Liu, Wei, A.D.E.D.S.C.R.S.F.C.Y., Berg, A.C., 2016. Ssd: Single shot multibox detector, in: Leibe, Bastian, M.J.S.N.a.W.M. (Ed.), Computer Vision – ECCV 2016, Springer International Publishing, Cham. pp. 21–37.

Lucas, Christian, S.J.J.K.A.A.L.F., Heinrich, M.P., 2019. Automatic detection and segmentation of the acute vessel thrombus in cerebral ct, in: Handels, Heinz, D.T.M.M.A.M.H.K.H.P.C.T.T. (Ed.), Bildverarbeitung für die Medizin 2019, Springer Fachmedien Wiesbaden, Wiesbaden. pp. 74–79.

Martins-Filho, R.K., 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. doi:10.1016/j.jstrokecerebrovasdis.2018.12.003.

Mayer, Viarasilpa T., P.N.B.M.e.a., 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. doi:10.1161/STROKEAHA.119.027356.

Meijs, Meijer FJA, P.M.G.B.M.R., 2020. Image-level detection of arterial occlusions in 4d-cta of acute stroke patients using deep learning. Med Image Anal 66, 101810. doi:10.1016/j.media.2020.101810.

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. doi:10.3390/diagnostics12030698.

Monkam, P., Qi, S., Ma, H., Gao, W., Yao, Y., Qian, W., 2019. Detection and classification of pulmonary nodules using convolutional neural networks: A survey. IEEE Access 7, 78075–78091.

Murray, Unberath M, H.G.H.F., 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.

Paola Martinez Arias, Uma Maria Lal-Trehan Estrada, M.T.L.G.A.O.X.L., 2023. Binary classification and detection of large-vessel occlusions in acute ischemic stroke. MAIA Erasmus Mundus Intake 2021-23. URL: https://maiamaster.udg.edu/master-thesis-proceedings/.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788. doi:10.1109/CVPR.2016.91.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks, in: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.

Shafaat, S.H., 2023. Stroke imaging. Treasure Island (FL): StatPearls Publishing .

Stib, Vasquez J, D.M.e.a., 2020. Detecting large vessel occlusion at multiphase ct angiography by using a deep convolutional neural network. Radiology 297(3), 640–649. doi:10.1148/radiol.2020200334.

Sweid, Hammoud B, R.S.e.a., 2019. Acute ischaemic stroke interventions: large vessel occlusion and beyond. Stroke and Vascular Neurology 5, 80–85. doi:10.1136/svn-2019-000262.

Tang, H., Kim, D.R., Xie, X., 2018. Automated pulmonary nodule detection using 3d deep convolutional neural networks. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018) , 523–526.

Tolhuisen, M.L., Ponomareva, E., Boers, A.M.M., Jansen, I.G.H., Koopman, M.S., Sales Barros, R., Berkhemer, O.A., van Zwam, W.H., van der Lugt, A., Majoie, C.B.L.M., Marquering, H.A., 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. doi:10.3390/app10144861.

Tsao, C.W., Aday, A.W., Almarzooq, Z.I., Anderson, C.A., Arora, P., Avery, C.L., Baker-Smith, C.M., Beaton, A.Z., Boehme, A.K., Buxton, A.E., Commodore-Mensah, Y., Elkind, M.S., Evenson, K.R., Eze-Nliam, C., Fugar, S., Generoso, G., Heard, D.G., Hiremath, S., Ho, J.E., Kalani, R., Kazi, D.S., Ko, D., Levine, D.A., Liu, J., Ma, J., Magnani, J.W., Michos, E.D., Mussolino, M.E., Navaneethan, S.D., Parikh, N.I., Poudel, R., Rezk-Hanna, M., Roth, G.A., Shah, N.S., St-Onge, M.P., Thacker, E.L., Virani, S.S., Voeks, J.H., Wang, N.Y., Wong, N.D., Wong, S.S., Yaffe, K., Martin, S.S., null null, 2023. Circulation 147, e93–e621. doi:10.1161/CIR.0000000000001123.

Xie, H., Yang, D., Sun, N., Chen, Z., Zhang, Y., 2019. Automated pulmonary nodule detection in ct images using deep convolutional neural networks. Pattern Recognit. 85, 109–119.

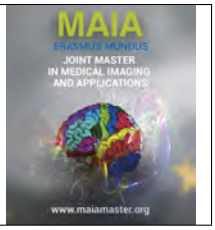
Zhou, Y., Tuzel, O., 2017. Voxelnet: End-to-end learning for point cloud based 3d object detection. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition , 4490–4499.

Zhu, W., Liu, C., Fan, W., Xie, X., 2018. Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) , 673–681.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3d u-net: Learning dense volumetric segmentation from sparse annotation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention. URL: https://api.semanticscholar.org/CorpusID:2164893.

Medical Imaging and Applications

Master Thesis, June 2024




# NeuroSculpt: Forecasting Brain Structure 9 Years Ahead Using Structural MRI

Agustin CARTAYA, Asta Håberg

*Department of Neuromedicine and Movement Science - NTNU, Trondheim, Norway*

**Abstract**

As people age, their brains undergo various structural transformations, primarily involving tissue loss. Accelerated changes can lead to serious conditions such as dementia or Parkinson's disease. Early detection of such abnormal changes in healthy individuals is crucial, as it may allow for early interventions to mitigate these consequences. However, continuous Magnetic Resonance Imaging (MRI) studies, necessary for such detection, are both time-intensive and costly. Currently, several alternatives have been proposed to predict brain structural changes using advances in machine learning and deep learning. However, most focus on patients with neurodegenerative diseases and none specialize in healthy adult populations. In this study, we aimed to predict structural brain changes over a span of nine years in a healthy adult population. We used 3D T1-weighted MR images and explored two primary family of methods. The first family was based on Deformation Fields (DFs), while the second employed deep learning techniques using Generative Adversarial Networks (GANs). DF-based methods were built on the hypothesis, that brain changes observed in one subset of individuals could predict changes in others within the same population. The GAN-based methods were inspired by advancements in predicting brain changes in infants and Alzheimer's disease patients. We evaluated the results of these methods using various assessment criteria, including image similarity, similarity of brain regions, and total brain atrophy. Our results indicated that DF-based techniques were more effective and stable than GANs, demonstrating a greater ability to capture subtle changes, particularly in the thalamus and cortex, as well as significant changes in the ventricles in line with our hypothesis. In contrast, GAN-based methods primarily predicted volumetric changes in the ventricles. This study provided a foundation for future research in brain change prediction, highlighting the effectiveness of DF-based methods and suggesting improvements for GAN approaches.

*Keywords:* Brain Aging, Deformation Fields, GANs

## 1. Introduction

### 1.1. Longitudinal Prediction

Longitudinal prediction involves anticipating how certain characteristics of an individual will change over time based on data collected at earlier moments or theoretical models that describe possible patterns of change (Caruana et al., 2015). In neurology, this approach is crucial for forecasting the progression of neurodegenerative diseases such as Alzheimer's, Parkinson's, or Multiple Sclerosis, enabling treatments before clinical symptoms become evident and slowing disease progression (Arya et al., 2023; Coll et al., 2023; Li et al., 2019). However, despite its benefits, longitudinal prediction faces several challenges, as the accuracy of predictions heavily depends on the quality and quantity of available data, which is not always easy to obtain, especially in the medical domain. (Bandettini, 2012; Bernal et al., 2021; Modat et al., 2014).

### 1.2. Brain Changes with Aging

As the brain ages, significant structural and functional changes occur that primarily affect cognition (Schulz et al., 2022). On a large scale, grey matter (GM) and white matter (WM), which contain neuronal cell bodies and long-distance synapses, respectively, undergo atrophy, being replaced by cerebrospinal fluid (CSF) (Ge et al., 2002). Some studies indicate that certain brain structures are more susceptible to aging-related changes (Choi et al., 2022; Fujita et al., 2023; Raz et al., 2005).

Structures such as the hippocampus, thalamus, and cortex, crucial for memory, sensory information transmission, and complex cognitive functions, show significant atrophy. These changes are reflected in the expansion of the ventricles, which dilate to compensate for brain volume loss, and the increase of CSF around the brain due to the reduction in the height between sulci and gyri (Kaye et al., 1992).

Brain aging varies between healthy individuals and those with neurodegenerative diseases (Habes et al., 2016). In healthy individuals, structural changes are generally slower and more subtle, influenced by genetics and lifestyle (Mulugeta et al., 2022). In contrast, in patients with diseases like Alzheimer's, atrophy is more accelerated and follows specific, well-documented patterns (Pini et al., 2016). Consequently, numerous predictive models for neurodegenerative diseases have been developed (Arya et al., 2023).

To study these age-related brain changes, Magnetic Resonance Imaging (MRI) has been used as a fundamental tool due to its ability to provide detailed visualization of brain structures (Vemuri et al., 2015). Particularly, T1-weighted (T1w) MR images are especially useful for anatomical visualization, offering good resolution and contrast between GM, WM and CSF (Chen et al., 2018). These images also allow the observation of the subcortical structures sensitive to aging (Duan et al., 2020), thereby facilitating the monitoring of structural changes associated with aging and neurodegenerative diseases.

### 1.3. Methods for Longitudinal Brain Prediction

Over the last decades, advances in machine learning have offered a powerful tool in longitudinal neurological studies, allowing the quantification of brain aging in patients with neurodegenerative diseases (Zapaishchykova et al., 2024). Currently, two families of methods are most commonly used to infer longitudinal brain changes:

- The first and most used is based on Deformation Fields (DFs). A DF is a fundamental element in the area of non-rigid registrations (Crum et al., 2004) and is based on a vector field that indicates how each pixel (or voxel in 3D images) of a moving image M should be displaced to align it with a fixed image F.
- The second, more recent and based on advances in deep learning, uses generative adversarial networks (GANs) (Goodfellow et al., 2014). A GAN consists of two neural networks: a Generator that creates images from an input and a Discriminator that evaluates their realism, competing with each other to continuously improve.

In the context of longitudinal brain prediction, methods of the first family seek to infer a DF that explains structural changes over time, which can then be applied to the initial brain images to obtain their evolution using image registration. Meanwhile, methods in the second family train a GAN for image-to-image translation (Isola et al., 2018a) using historical data (e.g., initial and future images), and then predict the brain's evolution given the initial image.

### 1.4. Predicting Brain Changes in Healthy Populations

While the majority of research focuses on structural brain changes caused by neurodegenerative diseases (Camara et al., 2006; Rachmadi et al., 2019; Ravi et al., 2019; Xia et al., 2021), there is significant value in extending these predictive models to healthy populations. Predictive models tailored for healthy individuals could offer insights into normal aging trajectories, identify atypical changes indicative of early disease onset, and highlight the impact of lifestyle and genetic factors on brain health (Hedman et al., 2012). Moreover, such models could facilitate early interventions, potentially mitigating the risk of developing neurodegenerative conditions (Rachmadi et al., 2019). However, predicting brain changes in healthy populations presents challenges, such as the variability of aging processes due to the influence of individual's sociodemographic, health, genetics and lifestyle factors (Mulugeta et al., 2022) and the need for extensive longitudinal data (Bethlehem et al., 2021). This naturally leads to the question: Is it possible to predict brain changes in healthy populations?

### 1.5. Objective of the Master's Thesis

The objective of this project is to address the previous question and, specifically, to attempt to predict structural brain changes over a nine-year period in healthy adults using 3D T1w MR images. with participants having an average age of 60 years at the time of the initial scan (baseline) and 69 years at the time of the second scan (follow-up).

To achieve our objective, we implemented various methods based on the two main families of longitudinal prediction mentioned earlier:

- **DF-Based Methods:** These methods are based on inferring a DF that captures the necessary volumetric changes to register the baseline and, consequently, predict the follow-up scan. First, we create a dataset of deformation atlases by registering baseline to follow-up and obtaining the resulting DFs from a subset of our population. Then, we implement four different methods based on variants of multi-atlas techniques (Iglesias and Sabuncu, 2014) to combine the obtained deformation atlases and create the desired DF.

- **GAN-Based Methods:** In this family, the methods are based on training a GAN with baseline and follow-up scans from a subset of our population, allowing it to learn the longitudinal changes. Then, from the baseline of a new individual, the GAN can predict the follow-up. To achieve this, we implemented four different GANs based on the architectures proposed by Peng et al. (2021), Huang et al. (2022) and Choi et al. (2020) and adapted them to our objective.

Finally, we conducted a statistical analysis to determine the best method of each family and overall. We used various comparison metrics between the predicted and expected images, based on image similarity, similarity of brain structures relevant to aging (Choi et al., 2022; Fujita et al., 2023), and total brain atrophy using the Brain Parenchymal Fraction (BPF) (Rudick et al., 1999).

## 2. State of the art

During our review of the state of the art, our primary focus was on longitudinal brain changes, where most of the works we found employed DF-based techniques or GAN-based techniques, primarily for predicting brain atrophy. Additionally, we expanded our search to facial aging studies as they also presented innovative techniques in longitudinal prediction.

### 2.1. DF-Based Approaches

The prediction of brain atrophy in patients with Alzheimer's or other neurodegenerative diseases has been extensively researched in recent years, primarily using models that aim to infer a DF with specific volumetric changes. Smith et al. (2003) presented a biomechanical model using finite element method and applied thermal loads to induce expansion or contraction in the desired tissues by a DF. Camara et al. (2006) expanded this approach with a thermoelastic model and added acquisition artifacts to the generated image for greater realism. Karacali and Davatzikos (2006) and Sharma et al. (2010) presented models that minimize an energy function, penalizing the deviation between the desired volumetric loss and that inferred from the Jacobian of the DF, preserving brain topology and allowing free movement of CSF. Modat et al. (2014) employed multimodal registrations to obtain a set of velocity fields describing actual brain changes, subsequently combining them to generate DFs specific to each type of disease. Khanal et al. (2017, 2016) developed a biophysical model to generate a DF based on Stokes equations from fluid mechanics, but with a non-zero mass source term to allow the deformation of each tissue based on its prescribed atrophy. Da Silva et al. (2020) used deep neural networks to predict the DF from an atrophy map. In a subsequent work, Da Silva et al. (2021) presented a more comprehensive model that infers the atrophy map from the patient's medical data. More recently, Bernal et al. (2021) proposed a cascade U-Net (Ronneberger et al., 2015) approach to generate controlled synthetic volumes based on probability maps of altered tissues.

Many of these methods propose quite accurate prediction results. However, except for Modat et al. (2014) and Da Silva et al. (2021), these results depend on prespecified atrophy maps. This reliance can be limiting because intermediary scans between baseline and follow-up are needed to construct these maps and observe specific changes for each patient. Given that our dataset does not contain intermediary scans, we propose DF-based models that infer changes based on inter-individual similarity rather than relying on atrophy maps.

### 2.2. GAN-Based Approaches

Recent research using GANs has demonstrated their utility in predicting the progression of neurodegenerative diseases and aging in MRIs. Rachmadi et al. (2019) proposed DEP-GAN to predict the evolution of white matter hyperintensities in patients with small vessel disease. This model combines GAN with Irregularity Maps to generate Disease Evolution Maps. Similarly, Ravi et al. (2019) and Xia et al. (2021) presented models to predict the evolution of atrophy in brain MRI as a function of age and Alzheimer's disease status. The former proposed DaniNet, a model that combines a conditional deep autoencoder with a GAN, integrating biological constraints to predict realistic synthetic images. The latter developed a network that does not require longitudinal data for training, using identity-preserving losses to maintain subject-specific features in the predicted images. More recently, Gadewar et al. (2023) employed a style-transfer-based architecture to predict brain changes in subjects aged 60 to 79, using multiple age and sex-specific domains. In the field of infant brain development, Peng et al. (2021) and Huang et al. (2022) focused on longitudinal prediction of structural and contrast changes in infants over the first year of life. The first work introduced MPGAN, which combines a feature extractor with a GAN to generate high-quality images using perceptual loss. The second work addressed the problem differently with MGAN, a GAN-based network that uses spatial and frequency information from the baseline to predict metamorphic changes.

All these approaches underscore the capability of GANs for predicting brain changes, but they present several limitations. First, training with 2D slices (Gadewar et al., 2023; Rachmadi et al., 2019; Ravi et al., 2019; Xia et al., 2021), which in most cases is not a choice but rather unavoidable due to lack of computational resources, may result in the loss of inherent 3D information in structural MRI. Second, although training

without longitudinal data is innovative (Gadewar et al., 2023; Xia et al., 2021), it lacks mechanisms to verify the results and guide the network toward individual-specific predictions. Finally, most of the presented works validate their results using global image metrics, which do not detect subtle structural brain changes, mainly in subcortical regions, which are important in brain aging.

In three of our proposed GAN-based models, we address the challenge of loss of 3D information by employing 3D models and reducing image bit-depth to conserve memory. We overcame the second challenge by leveraging our dataset's longitudinal images. We meticulously evaluate model performance and guide training through tailored loss functions designed for individualized longitudinal changes. Furthermore, we present results specific to different brain regions and evaluate them using different metrics.

### 2.3. Facial Aging

Studies on facial aging propose a different and innovative approach that can be adapted for longitudinal brain prediction, as demonstrated by Ravi et al. (2019) and Gadewar et al. (2023). Among the most notable methods found are those by Antipov et al. (2017) and Choi et al. (2020), which propose GAN-based models. The former proposed Age-cGAN, which generates aged images while preserving the individual's identity. The process uses an encoder to find an optimal latent vector allowing the generator to reconstruct the image; then, the age category in the generator's input is changed to produce the image with the desired age. To ensure identity preservation, a pretrained facial recognition network is used. In the second method, they proposed StarGAN-v2, a network that can transform images from one domain to another with diversity and variability. It implements a style encoder that extracts features (e.g., hairstyle and facial characteristics) from an image A and a generator that adds those features to an image B. Some more recent works implemented diffusion models (Sohl-Dickstein et al., 2015). In Chen and Lathuilière (2023), they used a model that inverts the input image to a latent noise and performs local age-guided text and attention control editing to achieve precise and realistic transformations. In another method proposed by Banerjee et al. (2023), a latent diffusion model with contrastive and biometric losses is used, preserving identity and achieving realistic and high-fidelity age modifications.

These approaches offer different sources of inspiration for longitudinal prediction. However, all these methods rely on 2D images and must be adapted to work with 3D MRI scans, which could be challenging due to misaligned slices. To overcome this limitation in our fourth GAN-based model, we ensure accurate alignment between baseline and follow-up during preprocessing. Additionally, we implemented a dataloader capable of handling inter-individual slice alignment.

## 3. Material and methods

### 3.1. Data

In our study, we used a total of 703 individuals from the Nord-Trøndelag Health Study (HUNT) (Åsvold et al., 2022), a longitudinal study involving a healthy population from Nord-Trøndelag, Norway, since 1984. Our study focuses solely on using the 3D T1w MR images obtained during the third wave (HUNT3) (Håberg et al., 2016) in 2009 to predict images from the fourth wave (HUNT4) collected in 2018. The HUNT3 images were obtained using a 1.5T General Electric scanner with a resolution of $1.25 \times 1.25 \times 1.20\,\mathrm{mm}^3$, while the HUNT4 images were acquired using a 3T General Electric scanner with an isotropic resolution of 1 mm. Appendix A provides more information about HUNT3 and HUNT4 T1w MR scans. In this study, we randomly divided the dataset into two main sets for training and testing, with 620 and 83 individuals respectively. Depending on the method employed, validation subsets were also taken from the training set.

### 3.2. Preprocessing

Given that the baseline and follow-up were obtained nine years apart and with different magnetic field strengths, we harmonized the whole dataset applying a preprocessing. This was performed using FreeSurfer tools (Fischl, 2012) and its deep learning implementation FastSurfer (Henschel et al., 2020).
We began the preprocessing by converting the images to 1mm isotropic MP-RAGE format using the SyntSR tool. This was done for both HUNT3 and HUNT4 images, as employing this network also facilitated bias field correction and contrast standardization, as indicated in the original work (Iglesias et al., 2023, 2021). Then, we aligned the individuals to the MNI-ICBM 152 2009c space (Fonov et al., 2011, 2009) using affine registration with mri_robust_register (Reuter et al., 2010). To ensure that each individual's baseline was adequately aligned with their follow-up, we first registered the baseline to the MNI space and then registered the follow-up to its corresponding registered baseline scan. Finally, we performed skull stripping using SynthStrip (Hoopes et al., 2022), followed by normalization to extract only the brain region within an intensity range of $[0, 1]$. During preprocessing, we obtained two brain masks, with and without the cerebellum, and three types of tissue segmentation. The first segmentation included the 3 primary tissues: CSF, GM, and WM. The second segmentation delineated 35 tissues, incorporating subcortical structures, while the third segmentation encompassed 95 tissues, including both subcortical structures and various cortical regions. The final size in voxels of

the resulting baseline and follow-up images, along with their segmentations and masks, was $193 \times 229 \times 193$. Figure 1 shows the complete preprocessing pipeline and the results of the obtained images.

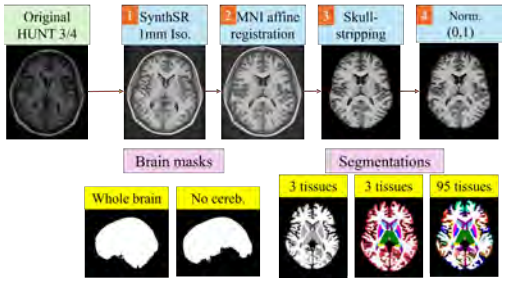



Figure 1: **Preprocessing Pipeline:** Steps performed during preprocessing and the obtained brain masks and segmentations.

### Notation and Main Objective

Hereafter, we will refer to the training set for the baseline scans as $TX_0$ and for the follow-up scans as $TX_1$, while the test set is referred to as $X_0$ for the baseline and $X_1$ for the follow-up scans. Our primary objective is to find $\hat{x}_1$, the best possible approximation of $x_1 \in X_1$, based on the corresponding baseline $x_0 \in X_0$. To achieve this, we employed several methods derived from the two main families of longitudinal brain prediction, which are detailed in the subsequent sections.

### 3.3. DF-Based Methods

**Hypothesis —** Our primary hypothesis for this family of methods is that the brain changes of an individual from a specific population could be predicted using the brain changes of other individuals from the same population.

The first step to verify our hypothesis was an inter-individual statistical analysis. We evaluated the similarity in both baseline and follow-up scans to determine if individuals with similar brain structures at baseline maintained this similarity at follow-up in our dataset. For each individual $I_0$, we identified the individual $I_1$ with the highest baseline similarity to $I_0$, and checked if $I_1$ remained the most similar to $I_0$ at follow-up or was among the top N most similar individuals. Table 1 and Fig. 2 shows the results of this analysis. To compute the similarity between individuals, we tested two metrics: the Structural Similarity Index (SSIM) introduced by Wang et al. (2004), which ranges from -1 to 1, where 1 indicates perfect similarity, 0 indicates no similarity, and -1 indicates perfect anti-correlation; and the mean Dice coefficient across the three main tissues, which ranges from 0 to 1, where 1 indicates perfect overlap. For two tissues $A$ and $B$, the Dice coefficient is defined as follows:

$$dice = \frac{2|A \cap B|}{|A| + |B|}$$

We tested these two metrics to capture different aspects of brain structure similarity. SSIM provides a global assessment of structural information and visual quality, while the Dice coefficient focuses on tissue correspondence.

Table 1: Inter individual similarity consistency (%)

| Metric | MS | Top3 | Top5 | Top10 | Top15 |
|---|---|---|---|---|---|
| SSIM | 67 | 93 | 97 | 99 | 100 |
| Mean $dice_3$ | 58 | 82 | 91 | 96 | 99 |

Probability that $I_1$ is the most similar (MS) to $I_0$ at follow-up or is among the topN most similar individuals using different similarity metrics.

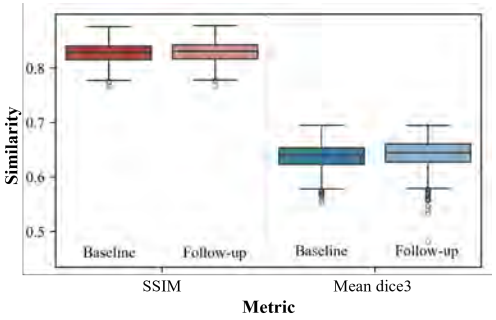



Figure 2: **Inter-Individual Similarities.** Similarity between $I_0$ and $I_1$ at baseline and follow-up for all individuals using different metrics.

Results in Table 1 demonstrated that $I_1$ was consistently identified as the most similar individual to $I_0$ at follow-up with a probability of 67% using SSIM and 58% using the mean Dice coefficient. Additionally, $I_1$ was among the top5 with over 90% probability using both metrics. Furthermore, as shown in Figure 2 the similarity between $I_0$ and $I_1$ remained stable from baseline to follow-up. These results confirmed that individuals with similar brain structures at baseline maintained this similarity at follow-up in our dataset and motivated us to proceed with the second part of the hypothesis evaluation.

For this second part, we obtained a dataset of DFs from the training data, which we called $TDf$ with $tdf$ as one of its elements. This was achieved by applying non-rigid registrations to the images of $TX_0$ towards their corresponding images in $TX_1$ using Elastix (Klein et al., 2009). For these registrations, we used B-Spline transformations with advanced normalized cross-correlation as the similarity metric and a pyramidal approach. Additional information about the registration can be found in Appendix B.

Next, we calculated an average DF from $n$ $tdf_i$, $i \in [1, n]$, and used it to register the images of $X_0$. The obtained registered scans indicated an improvement in all

individuals compared to the initial differences between the baseline and follow-up scans. The results showed a mean improvement of 4.1% for the Dice coefficient of the CSF, as well as 0.7% and 0.5% for the GM and WM, respectively. The image similarity based on the SSIM also improved by 0.8%. More details about these results can be found in the Results Section 4.2. These findings confirmed that the use of an average DF, based on a subset of a population, can effectively infer some brain changes in the remaining population, corroborating our initial hypothesis. This prompted us to develop our DF-based methods explained in the following sections.

**Objective —** Our objective with the following four methods is to infer $\hat{df}$, a DF that explains the longitudinal volumetric changes, allowing us to register $x_0$ to obtain $\hat{x}_1$. We base these methods on multi-atlas techniques and an adaptation of the K-Nearest Neighbors algorithm to combine the elements of $TDf$ and obtain $\hat{df}$.

#### 3.3.1. Similar Images

Here, we attempted to use image similarity to infer $\hat{df}$. Initially, we calculated the similarity $s$ between $x_0$ and each $tx_0 \in TX_0$, selecting the $n$ most similar $tx_{0i}$ with their corresponding $tdf_i$, $i \in [1, n]$. Subsequently, we weighted the $tdf_i$ with their respective normalized $s_i$ and computed their average, resulting in $\hat{df}$ (see Equation (1)). In the implementation of the method, we used L1 normalization. Additionally, we tested different similarity metrics and values for $n$ to evaluate their impact on the final predictions.

$$\hat{df} = \frac{\sum_{i=1}^{n} s_i \cdot tdf_i}{\sum_{i=1}^{n} s_i} \tag{1}$$

#### 3.3.2. Similar Images with Registration

In this method, we followed a similar approach to the previous one, but with one key difference: after identifying the $n$ $tdf_i$, we registered them to the $x_0$ space before computing the weighted average (see Equation (2)). We adopted this approach because we considered that obtaining a more precise alignment of the starting point of each vector from a given $tdf_i$ with respect to the image $x_0$ might result in a more accurate deformation of certain tissues. To register a $tdf_i$ to the $x_0$ space and obtain $tdf_{i_{x0}}$, we first applied a registration of $tx_i$ to $x_0$ to obtain the necessary deformation, and subsequently applied it to the corresponding $tdf_i$. All registrations were made using Elastix, and we tested two different types of registration, affine and B-spline.

$$\hat{df} = \frac{\sum_{i=1}^{n} s_i \cdot tdf_{i_{x0}}}{\sum_{i=1}^{n} s_i} \tag{2}$$

#### 3.3.3. Similar Patches

In this approach, we aimed to infer $\hat{df}$ by patches to capture more anatomical variability. First, we obtained $m$ overlapping uniform patches $p$ of size $w$ that covered the entire $x_0$. Similarly, we proceeded with all $tx_0$ and their corresponding $tdf$, generating $tp$ and $tdfp$ respectively. Then, given a patch $p_j$, $j \in [1, m]$, we calculated the similarity $s$ between $p_j$ and each $tp_j$. Next, we selected the $n$ most similar $tp_j$ and finally we computed the weighted average of their corresponding $tdfp_j$ to obtain $\hat{dfp}_j \in \hat{df}$. This process was repeated for each $p_j$ to reconstruct the complete $\hat{df}$ (see Equation (3)). During reconstruction, we used a spline-based method to address overlapping, which helped minimize artifacts in the overlapping areas. In this approach, we set $w = 32$ and an overlap of 50%, both values were experimentally favorable. During the evaluation of the method, we used different values of $k$ and $n$ to assess their effects on the final prediction.

$$\hat{df} = \bigoplus_{j=1}^{m} \hat{dfp}_j \tag{3}$$

Where $\bigoplus$ denotes the operation of patch concatenation with overlap, and each $\hat{dfp}_j$ is constructed as follows:

$$\hat{dfp}_j = \frac{\sum_{i=1}^{n} s_{ji} \cdot tdfp_{ji}}{\sum_{i=1}^{n} s_{ji}}$$

The similarity metric used between patches is based on a weighted Dice coefficient with $k$ tissues, as explained in the following equation:

$$s_j = \frac{\sum_{q=1}^{k} (w_q + a1_q + a2_q) \cdot dice_q(p_j, tp_j)}{2 + \sum_{q=1}^{k} w_q} \tag{4}$$

Where $\text{dice}_q(x, y)$ is the Dice coefficient for tissue $q$ between the segmentation with $k$ tissues of $x$ and $y$; $a1_q$ and $a2_q$ are the areas of tissue $q$ with respect to the patch size; and $w_q$ is a weigh given to each tissue.

#### 3.3.4. Similar Tissues

Here, we aimed to reconstruct $\hat{df}$ by tissues to allow variability and ensure that each individual tissue deforms consistently. To do this, we used the segmentation with $k$ tissues $segk_0$ of $x_0$ as well as the segmentations $tsegk_0$ of $tx_0$ and reconstructed a unique DF for each tissue, subsequently combining them to form $\hat{df}$. This was done very similarly to the patch approach 3.3.3 but with tissue regions instead of patches (see Equation (5)). In this case, there was no overlapping since $segk_0$ contains mutually exclusive tissues. The used similarity metric between the tissues was the Dice coefficient, and during the evaluation, we used different values of $k$ and $n$.

$$\hat{df} = \bigcup_{j=1}^{k} \hat{dfseg}_j \tag{5}$$

Where $\bigcup$ denotes the operation of tissue concatenation and each $\hat{dfseg}_j$ is constructed as follows:

$$\hat{dfseg}_j = \frac{\sum_{i=1}^{n} s_i \cdot tdfseg_{ji}}{\sum_{i=1}^{n} s_i}$$

### 3.4. GANs-Based Methods

**Objective —** Our primary objective with the following four methods is to train a GAN to predict $TX_1$ from $TX_0$, enabling it to learn to infer longitudinal structural changes. This way, given an $x_0$, the network's generator can predict $\hat{x}_1$. To this end, we implemented the architectures proposed by Peng et al. (2021), Huang et al. (2022), and Choi et al. (2020) and adapted them to our objective. In the following methods, we refer to $\hat{tx}_1$ as a predicted image during training and $tx_1 \in TX_1$ as the expected image.

#### 3.4.1. MPGAN

In this approach, we used the multi-contrast perceptual adversarial network MPGAN proposed by Peng et al. (2021). Originally, this network was built to predict longitudinal changes in infant brains during the first year of life, which undergo quite different changes compared to adult aging brains (Huang et al., 2022). In the original paper, they proposed a simple architecture and a multimodal one; in our case, we only implemented the first one given our dataset.

**Network Architecture —** The MPGAN architecture consists of three main components: A Generator (G) using a U-Net architecture with residual blocks in both the encoder and decoder; a Discriminator (D) that is a classifier composed of convolutional layers followed by an output layer with sigmoid activation; and a pre-trained feature extractor ($\phi$) to extract perceptual features. To build $\phi$ they used the encoder part of the architectured proposed by (Zhou et al., 2019) which is a U-Net model trained with 3D medical images.

**Loss Functions —** The original paper proposed three loss functions: An adversarial loss ($L_{\text{adv}}$), original to GANs (Goodfellow et al., 2014), which helps $\hat{tx}_1$ approach the distribution of $TX_1$. A voxel-wise reconstruction loss ($L_{\text{vr}}$), as introduced in Isola et al. (2018b), which ensures consistency between $\hat{tx}_1$ and $tx_1$ by penalizing voxel-to-voxel differences with an L1 loss. Finally, a perceptual loss ($L_{\text{p}}$), which helps produce sharper and more detailed images by penalizing the difference between the features extracted from $\hat{tx}_1$ and $tx_1$ using $\phi$. The total loss function used is the following:

$$L_{\text{total}} = L_{\text{adv}} + \alpha L_{\text{vr}} + \beta L_{\text{p}} \tag{6}$$

**Implementation Details —** We used TensorFlow and built the proposed architecture from scratch following the instructions of the original paper, as a functional source code was not available. The Adam optimizer (Kingma and Ba, 2017) with an initial learning rate of 2e-4 was employed, and we applied a decay of 0.5 and a patience of 10 epochs based on the validation loss. The trade-off coefficients $\alpha$ and $\beta$ were both set to 25, as proposed in the original paper.

To train the network, we used 80% of the images from $TX_0$ for training and 20% for validation in each epoch. The training process was conducted for a total of 100 epochs with a batch size of 1, applying early stopping with a patience of 10 to avoid overfitting. In order to save GPU memory and train the model using the complete 3D volumes, we used TensorFlow Mixed Precision, which employs both 16-bit and 32-bit floating-point types during training.

#### 3.4.2. MPGAN + Segmentation Loss

Here, we used the same MPGAN network explained in the previous section 3.4.1, with the addition of a segmentation similarity constraint. This was done to increase the similarity of the three main brain tissues (CSF, GM, WM) between $\hat{tx}_1$ and $tx_1$. In this way, we ensured that global structures and specific tissue details remained consistent, improving the accuracy of segmentation and the structural quality of the generated images.

**Loss Functions —** To calculate the segmentation loss ($L_{\text{seg}}$), we used a dice-based loss between the segmentation with three tissues of $\hat{tx}_1$ and $tx_1$ as shown below:

$$\begin{aligned} L_{\text{seg}} = 1 - \frac{1}{3} \big(&\text{dice}_{\text{CSF}}(\hat{tx}_1, tx_1) \\ &+ \text{dice}_{\text{GM}}(\hat{tx}_1, tx_1) \\ &+ \text{dice}_{\text{WM}}(\hat{tx}_1, tx_1)\big) \end{aligned} \tag{7}$$

Where $\text{dice}_q(x, y)$ is the same as used in Equation (4). For $tx_1$, the segmentation with three tissues obtained during preprocessing was used. However, for $\hat{tx}_1$, we had to calculate the segmentation during training. To achieve this, we used a Gaussian Mixture Model with priors based on the mean and variance of the tissues from $tx_1$. This allowed to calculate a segmentation for CSF, GM, and WM quickly and easily, with the possibility of gradient propagation in the loss function. Finally, we modified the total loss function as follows:

$$L_{\text{total}} = L_{\text{adv}} + \alpha L_{\text{vr}} + \beta L_{\text{p}} + \gamma L_{\text{seg}} \tag{8}$$

**Implementation Details —** The implementation was similar to the one described in the previous section 3.4.1, with the only difference being that we adjusted

$\alpha$, $\beta$, and $\gamma$ to 25, 20, and 15 respectively. These values were found to provide the best results for the validation set.

### 3.4.3. MGAN

For this method, we used the metamorphic generative adversarial network (MGAN) proposed by Huang et al. (2022). Similar to Peng et al. (2021), the original objective was to predict longitudinal changes in infant brains during the first year of life. However, in this work a 3D patch-based approach using spatial and frequency domains to capture metamorphic changes is proposed.

**Network Architecture —** The MGAN architecture is based on a CycleGAN (Zhu et al., 2020) and consists of two generators and two discriminators. Each generator includes an encoder, a spatial-frequency transfer block (SFT), and a decoder. The SFT is a dual-branch structure that captures and transforms information in both spatial and frequency domains. For the spatial domain, residual modules in series are used, and for the frequency domain, a discrete wavelet transform (DWT) is applied, followed by residual modules in series and finally an inverse DWT. This allows the preservation of structural and contrast details of the tissues throughout the reconstruction. On the other hand, the discriminators have a U-shaped architecture and generate voxel-level quality probability maps, guiding the generators to focus on the most challenging regions. Both the discriminators and generators use deep supervision in the decoder to strengthen the gradient flow and promote the learning of useful representations at multiple scales (Karnewar and Wang, 2020). It is worth noting that due to the cyclical nature of the network, it would also be possible to predict the baseline from the follow-up, but we did not use this functionality.

**Loss Functions —** The loss functions used in the paper include an adversarial loss ($L_{\text{adv}}$), a paired loss ($L_{\text{p}}$), and a cyclic loss ($L_{\text{cyc}}$) at different resolutions. The $L_{\text{adv}}$, has the same objective as explained earlier. The $L_{\text{p}}$ consists of several components: a quality loss ($L_Q$), which penalizes voxel-to-voxel differences with an L1 loss, using the discriminator results to focus on the more challenging regions to predict; a texture loss ($L_T$), which ensures that the texture of $\hat{tx}_1$ is similar to that of $tx_1$; and a frequency loss ($L_F$), which compares the wavelet representations between $\hat{tx}_1$ and $tx_1$ to preserve structural details. Finally, the cyclic loss ($L_{\text{cyc}}$), original to CycleGANs (Zhu et al., 2020), ensures cyclical consistency between the generated and real images, warranting that a transformed image, when reverted, is similar to the original. The total loss function implemented at each scale is the following:

$$L_{\text{total}} = L_{\text{adv}} + \alpha L_{\text{p}} + \beta L_{\text{cyc}} \tag{9}$$

Where:

$$L_{\text{p}} = L_Q + aL_T + bL_F \tag{10}$$

**Implementation Details —** We used TensorFlow and built the proposed architecture from scratch following the instructions of the original paper, as the source code was not available. The Adam optimizer (Kingma and Ba, 2017) with an initial learning rate of 1e-4 was employed, and we applied a decay of 0.5 and a patience of 10 epochs based on the validation loss. The trade-off coefficients for $\alpha$, $\beta$, $a$, and $b$ were set to 1, assuming these values were used in the original paper since they were not explicitly mentioned.

To train the network, we extracted patches of size $64 \times 64 \times 64$ with 50% overlap from the images of $TX_0$. These patches were selected to contain at least 15% brain tissue to avoid creating background-biased generators. The training process was conducted for a total of 10,000 epochs with a batch size of 1, ensuring that all patches from 80% of $TX_0$ were used for training, while the remaining 20% were reserved for validation.

### 3.4.4. StyleGAN

In this method, we used the StarGAN-V2 network proposed by Choi et al. (2020). This network was originally designed for style transfer between multiple domains with diversity in the generated images using a single Generator. In our case, we adapted the network similar to the work of Gadewar et al. (2023), to predict $\hat{x}_1$ from $x_0$ and a desired style $s$, taken from an element of $TX_1$.

**Network Architecture —** The StarGAN v2 architecture is based on four main elements: a generator (G), a mapping network (F), a style encoder (E), and a discriminator (D). G uses a U-Net-like architecture with an encoder, bottleneck, and decoder constructed with residual blocks. The style $s$ is injected into the decoder during the image reconstruction using adaptive instance normalization (Huang and Belongie, 2017). F is a multitask multilayer perceptron that generates a style code $s$ from a latent vector $z$ and a domain $y$. In our implementation, $z$ is a vector randomly sampled from a Normal Gaussian Distribution, and $y$ is an integer indicating whether the style belongs to the baseline or the follow-up. E is a multitask encoder that, given an image and its corresponding domain, extracts the style code $s$. Finally, D is a multitask discriminator that differentiates between real and generated images of a domain $y$. In this context multitask refers to the fact that the network has different output branches, one for each domain $y$. It is worth noting that all the networks were trained simultaneously. Due to the network's design, it is also possible to predict the baseline from the follow-up. However,

similar to the previous method, we will not focus on that functionality.

**Loss Functions —** The proposed loss functions consist of an adversarial loss($L_{adv}$) and a cyclic loss ($L_{cyc}$) with the same purpose as in the previous methods; a style reconstruction loss ($L_{sty}$) that forces the generator to use the style code $s$ when generating the image, extracting and comparing the style of $t\hat{x}_1$ with the desired style; and a style diversification loss ($L_{ds}$) that encourages the production of diverse images by regularizing the generator to explore different styles. The total loss function used is the following:

$$L_{total} = L_{adv} + \lambda_{cyc}L_{cyc} + \lambda_{sty}L_{sty} + \lambda_{ds}L_{ds} \quad (11)$$

**Implementation Details —** We used the code proposed by the original paper implemented in PyTorch and adapted it to our dataset. The training parameters we used were exactly the same as those proposed in the original paper.

To train the network, we used 2D slices extracted from the sagittal plane of $TX_0$ and $TX_1$. The 2D slices were extracted to contain at least 15% brain tissue to avoid creating a background-biased generator. The training process was conducted for a total of 100,000 epochs with a batch size of 4. It is worth mentioning that the dataloader we designed ensured that the network was trained with slices aligned among individuals.

### 3.5. Post-Processing

After obtaining the results, we applied post-processing to remove artifacts introduced during prediction, enhance overall image quality, and obtain brain masks and segmentations to evaluate the results. This was performed differently for both families:

- **DF-Based Methods:** We applied the brain mask and normalized the brain area to eliminate edge artifacts caused by interpolation during the registration. To obtain the brain masks and segmentations the initial segmentations and brain masks were registered with Elastix using the inferred DF.

- **GAN-Based Methods:** Here, we first processed the images with SynthSR to eliminate common GAN artifacts (Lee et al., 2023) and correct errors in image reconstruction from 3D patches (MGAN 3.4.3) or 2D slices (StyleGAN 3.4.4). Then, we performed skull stripping, followed by normalization in the brain area to remove the skull and background added by SynthSR. Finally, to obtain the brain masks and segmentations we used FastSurfer.

**Computational Resources**

For preprocessing and postprocessing, we used FreeSurfer installed on a Linux Ubuntu 18 PC with an Intel(R) Core(TM) i7-7700 CPU and 32GB of RAM, and FastSurfer Docker-version on a Windows 11 PC with a Intel(R) Core(TM) i9-12900H CPU, 32GB of RAM, and an NVIDIA GeForce RTX 3060 GPU with 6GB. For training deep learning methods, we utilized the High Performance Computing cluster at NTNU (IDUN). Specifically, we used clusters with NVIDIA V100 16GB GPUs for models trained using 2D slices and patches, and clusters with NVIDIA A100 40GB GPUs for models trained with full 3D volumes.

## 4. Results

In this section, we present the predicted scans obtained for each method using the test set. These predictions are evaluated with respect to the actual follow-up scans to verify their exactitude. To help the reader have a comprehensive overview of the evaluation we performed to choose the best method for each family and overall, we have structured this section in three main parts: First, we present the initial similarity between the baseline and follow-up scans of each individual and use it as the lower bound (LB), as it is expected that the results from the implemented methods will surpass this. Second, we present the results for each family separately and choose the best among them. Finally, we compare the best results from each family, conducting a more exhaustive analysis to decide the overall best method.

**Evaluation Metrics**

During the evaluation of the results, we used various comparison metrics based on global image similarity, cerebral tissue segmentation, and brain atrophy.

To choose the best method for each family, we used SSIM and the mean Dice coefficient of the three main tissues. This allowed us to quickly and accurately select the best results based on global structure and tissue correspondence.

For the more detailed analysis, we used the Dice coefficient, the Absolute Symmetrized Percent Volume Change (ASPVC), the Volume Fraction (VF), and the Brain Parenchymal Fraction (BPF). ASPVC has been used in other analyses of structural changes as it provides a dimensionless measure of variability between tissues (Khanal et al., 2016). For two tissues $A$ and $B$, ASPVC is defined as:

$$\text{ASPVC} = \frac{|A - B|}{0.5(A + B)} \cdot 100\%$$

On the other hand, VF helped us evaluate whether there was an increase or decrease in tissue volume. For

a tissue $A$, VP is defined as:

$$\text{VF} = \frac{A}{\text{Inter cranial volume}} \cdot 100\%$$

Finally, we used BPF to compute the atrophy inferred from the predicted scans, which is typically defined as the ratio of brain parenchymal volume to the intracranial volume. In our case, we computed the BPF using the GM and WM volumes ($V_{GM}$, $V_{WM}$), excluding cerebellum regions from the segmentation with three tissues, and the brain mask without cerebellum ($Bmask_{ncrb}$), as follows:

$$\text{BPF} = \frac{V_{GM} + V_{WM}}{Bmask_{ncrb}}$$

It is important to note that, for calculating both the BPF and the VF, we used the intracranial volume from the baseline scan to avoid potential segmentation errors. This approach is supported by extensive research demonstrating that total intracranial volume remains constant with aging (Brezova et al., 2014; Hansen et al., 2015; Pintzka et al., 2015).

### Significance Evaluation

To determine if the results of our methods were significantly different from the LB, we performed a paired t-test and we considered $p$-values below 0.01 to be statistically significant.

## 4.1. Initial Similarity

We started by evaluating the initial similarity between baseline and follow-up scans, setting this as LB for our methods. Table 2 and Figure 3 illustrate these initial similarities, providing a foundation for subsequent analyses.

Table 2: Initial Similarities Between Baseline and Follow-up Scans

| Initial | SSIM % ↑ | Mean dice % | | |
|---|---|---|---|---|
| | | CSF ↑ | GM ↑ | WM ↑ |
| **LB** | 94.6 ± 1.0 | 84.3 ± 4.8 | 80.5 ± 2.6 | 90.0 ± 1.8 |

Initial similarity metrics between baseline and follow-up scans, including SSIM and mean Dice coefficient for CSF, GM, and WM. The shown values are the mean of the test set, and the ± values represent the standard deviation. ↑ indicates that higher values are better.

## 4.2. Family-Wise Results

### 4.2.1. DF-Based Results

**Hypothesis results —** Before implementing the DF-based family of methods we evaluated our primarily hypothesis with different values for $n$ to verify its influence and modify this parameter in the actual methods. Table 3 shows the similarity of these results with the actual follow-up.

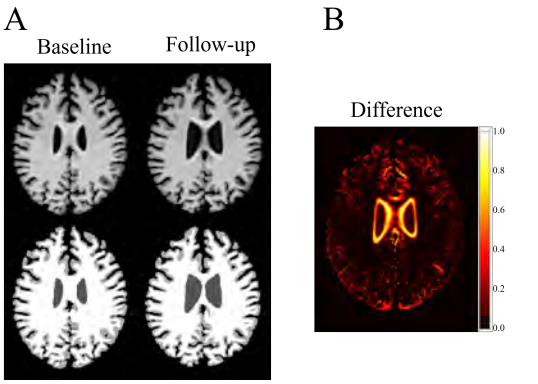



Figure 3: **Baseline and Follow-up Scans**. (A) shows the T1w scans in the first row and the segmentation of the three main tissues (CSF, GM, and WM) in the second row. (B) shows the difference image between the baseline and the follow-up; the lighter the color in a region, the more differences are present.

Table 3: Hypothesis Results - Similarities with Follow-up

| n | SSIM % ↑ | Mean dice% | | |
|---|---|---|---|---|
| | | CSF ↑ | GM ↑ | WM ↑ |
| 10 | 95.1 ± 0.8 | 87.8 ± 3.1 | *80.8 ± 2.1 | 90.4 ± 1.3 |
| 100 | 95.3 ± 0.8 | 88.4 ± 3.0 | 81.2 ± 2.3 | 90.5 ± 1.4 |
| 200 | 95.3 ± 0.8 | 88.3 ± 3.0 | 81.3 ± 2.3 | 90.5 ± 1.5 |
| 300 | 95.3 ± 0.8 | 88.3 ± 3.0 | 81.3 ± 2.3 | 90.5 ± 1.5 |
| 620 | 95.4 ± 0.8 | 88.3 ± 3.1 | 81.3 ± 2.3 | 90.5 ± 1.5 |

Similarity metrics between hypothesis predictions and actual follow-up scans using different values for $n$. * indicates $p$-values > 0.01.

The results obtained indicate a slight improvement between $n = 10$ and $n = 100$ but for $n > 100$, the changes are extremely small or negligible.

**DF-Based Methods Results —** For each method in this family, we evaluated different settings. For the Similar Images method 3.3.1, we tested two similarity metrics (SSIM and the mean Dice coefficient) and three values for $n = [5, 10, 100]$. In the Similar Images with Registration method 3.3.2, we evaluated two types of registrations (affine and non-rigid using B-Splines) and set $n = 5$. It is worth mentioning that the B-spline registration parameters were chosen to prioritize faster registration times over exhaustive optimization. For the Similar Patches method 3.3.3, we used different numbers of tissues $k = [3, 95]$ to evaluate similarity between patches and two values for $n = [10, 100]$. The hyperparameter $w$ in Equation 4 was set to 1 for all the tissues. Finally, for the Similar Tissues method 3.3.4, we tested different numbers of tissues to create the DF $k = [3, 95]$ and two values for $n = [10, 100]$. Table 4 shows the similarity with the follow-up for each method with their respective settings, and Figure 4 shows the predictions using the best setting for each method.

In this family of methods, all segmentation results were calculated using the registered segmentations. However, for the best method, the segmentation was re-

calculated from the predicted image using FastSurfer to avoid interpolation errors in discrete values caused by the registration. This result is also shown in Table 4 along with results obtained using ground truth deformations from the baseline to the follow-up scans through non-rigid registration with Elastix. This latter result could be interpreted as an upper bound (UB) for this family of methods.

Table 4: DF-Based Methods Results - Similarities with Follow-up

| Method | SSIM % ↑ | Mean dice % | | |
|---|---|---|---|---|
| | | CSF ↑ | GM ↑ | WM ↑ |
| **Images** | | | | |
| ssim 5 | 95.3 ± 0.7 | 89.7 ± 2.3 | 81.1 ± 2.1 | 90.7 ± 1.4 |
| ssim 10 | 95.4 ± 0.7 | 89.8 ± 2.4 | 81.5 ± 2.2 | 90.9 ± 1.4 |
| ssim 100 | 95.4 ± 0.8 | 89.6 ± 2.7 | 81.5 ± 2.3 | 90.9 ± 1.5 |
| dice 5 | 95.3 ± 0.7 | 89.6 ± 2.8 | 81.1 ± 2.1 | 90.7 ± 1.3 |
| dice 10 | 95.4 ± 0.7 | 89.8 ± 2.8 | 81.4 ± 2.1 | 90.9 ± 1.3 |
| -dice 100 | 95.4 ± 0.7 | 89.7 ± 2.8 | 81.6 ± 2.2 | 90.9 ± 1.3 |
| **Images Reg** | | | | |
| aff dice 5 | 95.3 ± 0.7 | 89.8 ± 2.9 | 81.2 ± 2.1 | 90.7 ± 1.3 |
| -bsp dice 5 | 95.4 ± 0.7 | 90.1 ± 3.0 | 81.8 ± 2.1 | 90.9 ± 1.3 |
| **Patches** | | | | |
| $seg_3$ 10 | 95.5 ± 0.7 | 90.4 ± 2.8 | 82.1 ± 2.2 | 91.2 ± 1.4 |
| $seg_3$ 100 | 95.5 ± 0.8 | 90.1 ± 2.9 | 82.1 ± 2.2 | 91.1 ± 1.4 |
| **-$seg_{96}$ 10** | **95.5 ± 0.7** | **90.5 ± 2.8** | **82.2 ± 2.2** | **91.2 ± 1.4** |
| $seg_{96}$ 100 | 95.5 ± 0.8 | 90.2 ± 2.9 | 82.1 ± 2.2 | 91.1 ± 1.4 |
| **Tissues** | | | | |
| $seg_3$ 10 | 95.5 ± 0.7 | 90.3 ± 2.5 | 81.6 ± 2.2 | 91.1 ± 1.4 |
| $seg_3$ 100 | 95.5 ± 0.8 | 90.0 ± 2.7 | 81.7 ± 2.3 | 91.1 ± 1.4 |
| -$seg_{96}$ 10 | 95.5 ± 0.7 | 90.4 ± 2.6 | 81.7 ± 2.3 | 91.1 ± 1.4 |
| $seg_{96}$ 100 | 95.5 ± 0.8 | 90.1 ± 2.7 | 81.7 ± 2.3 | 91.1 ± 1.5 |
| **Best post** | 95.5 ± 0.7 | 92.2 ± 2.8 | 84.1 ± 2.4 | 92.2 ± 1.5 |
| **UB** | 97.1 ± 0.3 | 94.8 ± 1.4 | 86.9 ± 0.7 | 93.7 ± 0.3 |

Similarity metrics between DF-based methods predictions and follow-up scans using different settings for each method. In the table, the methods are referred to as Images, Images Reg, Patches, and Tissues for Similar Images, Similar Images with Registration, Similar Patches, and Similar Tissues methods, respectively. '-' indicates the best method of each family, and the overall best method is indicated in **bold**. Best post and UB refer to the best method with the recalculated segmentation and the Upper Bound, respectively. ↑ higher is better.

As shown in Table 4 and Figure 4, all the results improved with respect to the LB and exhibit $p$-values $< 0.01$. The results of the methods are very similar when varying their hyperparameters. Despite this similarity, the patch-based and tissue-based methods show slight improvements over the others, particularly in CSF and GM for the patch-based method with 96 tissues and $n = 10$, which led us to select it as the best DF-based method.

### 4.2.2. GAN-Based Results

The next experiments we performed were using the GAN-based family. For the MPGAN 3.4.1 and MPGAN + Segmentation Loss 3.4.2 methods, we performed inference on the whole volume by feeding the network with the baseline scans. For the MGAN

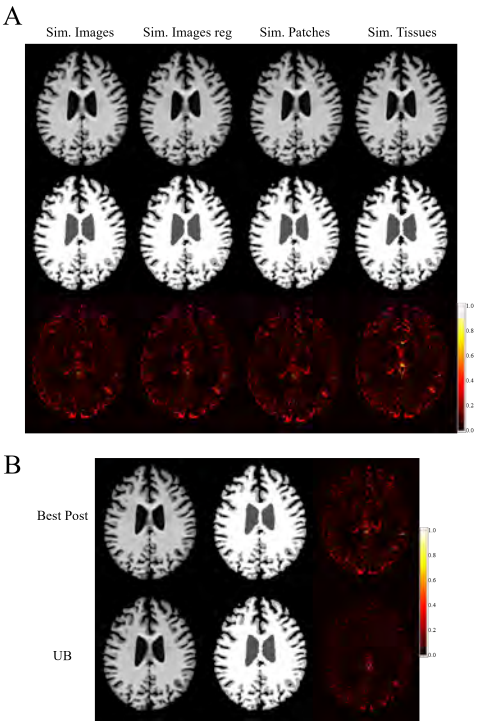



Figure 4: **DF-based Methods Predictions**. (A) Predictions of the DF-based methods using their best settings, including the segmentations of the three main tissues and the difference images with respect to the follow-up scans. (B) Prediction of the best method with the segmentation recalculated and prediction using ground truth deformation.

method 3.4.3, we extracted patches of size 64x64x64 with 50% overlap from the entire baseline scans, generating predictions for each patch. These patches were then assembled back together to create the complete volume. Similar to the method described in section 3.3.3, a spline-based method was used to handle the overlapping between patches and reduce the artifacts at the borders. Finally, for the StyleGAN method 3.4.4, since it required a style image to make the prediction, we selected the most similar baseline scan from our training set for each baseline scan in the test set and used its corresponding follow-up as the style. For each pair of images, we extracted slices from the sagittal plane and generated predictions for each slice. These slices were then stacked back together to reconstruct the complete volume.

After obtaining the predictions, we applied postprocessing to all the methods and then calculated the similarity results with the follow-up scans. These results are shown in Table 5, and the predicted images are displayed in Figure 5.

As shown in Table 5 and Figure 5, the predictions of most methods show results worse than the LB, with the MPGAN + Segmentation Loss method 3.4.2 being the only one that improved all metrics with a $p$-value $< 0.01$. Therefore, we selected it as the best GAN-based

Table 5: GAN-Based Methods Results - Similarities with Follow-up

| Method | SSIM % ↑ | Mean dice% | | |
|---|---|---|---|---|
| | | CSF ↑ | GM ↑ | WM ↑ |
| MPGAN | 92.1 ± 0.9 | 88.1 ± 3.1 | 72.2 ± 2.0 | 86.2 ± 1.6 |
| **MPGAN+seg** | **94.9 ± 0.7** | **90.7 ± 3.2** | **81.4 ± 2.2** | **91.0 ± 1.3** |
| MGAN | 92.6 ± 0.9 | 89.2 ± 2.9 | 72.7 ± 2.0 | 87.1 ± 1.5 |
| StyleGAN | 92.8 ± 0.8 | *82.4 ± 7.9 | 75.2 ± 1.8 | 87.1 ± 1.1 |

Similarity metrics between GAN-based methods predictions and follow-up scans. The best method is indicated in **bold**, and values lower than the LB are underlined. * indicates $p$-values > 0.01. ↑ higher is better.

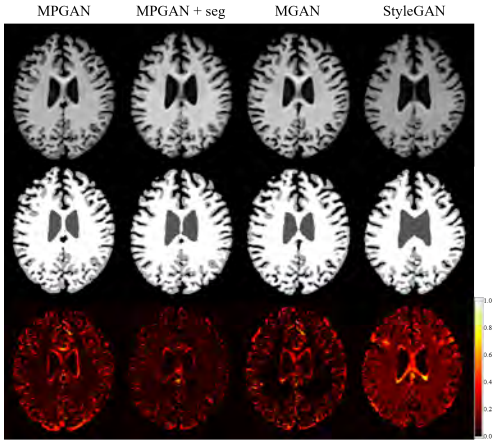



Figure 5: **GAN-based Methods Predictions.** Predicted T1w images, segmentations of the three main tissues, and difference images with respect to the follow-up scans.

method.

### 4.3. Best Methods Evaluation

**Tissue Based Analysis —** After selecting the best methods from each family, we conducted analyses based on cortical and subcortical structures to assess the ability of the predictions to capture subtle details. The structures selected for this analysis are presented in Figure 6. First, we assessed the volumetric changes of each structure using the VF to verify if the volumetric expansions or contractions were as expected. These results are shown in Table 6. Subsequently, we assessed the overlap and the volume differences between the structures from the predicted image and the actual follow-up using the Dice coefficient and ASPVC. These results are shown in Table 7 and 8.

**Atrophy Analysis —** We also performed an analysis based on the BPF to verify if the brain atrophy in the predicted images was similar to that in the actual follow-ups. During this analysis, we divided the test set into three groups with high, medium, and low BPF. This division allowed us to evaluate the predicted results for each group and verify the methods' performance. The results of this evaluation are presented in Figure 7.

**Visual Results —** Finally, we performed a visual inspection to determine if the computed metrics were consistent with the predicted scans. Figure 8 shows the predicted images of three individuals from each atrophy group. In this inspection, we also took into account the obtained segmentation highlighting the cortical and subcortical structures studied, as well as the difference image between the predictions and the actual follow-up.

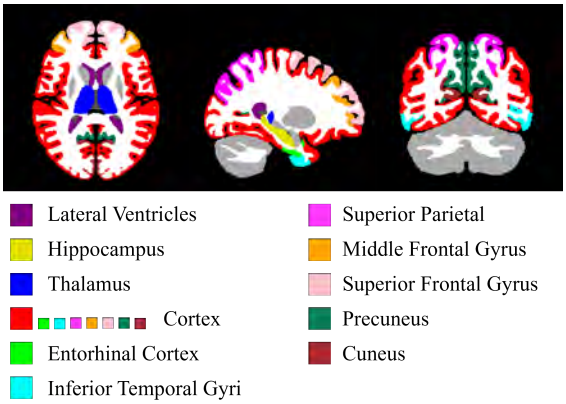



Figure 6: **Brain Structures Relevant to Brain Aging.** Eleven structures known to undergo marked changes with aging, used in this work to evaluate the accuracy of the predictions.

## 5. Discussion

Our research focused on determining the feasibility of predicting structural brain changes in healthy adults of around 60 years old over a nine-year period using 3D T1w MR images. We aimed to compare the accuracy of DF-based methods and GAN-based methods in predicting brain changes, evaluate their predictions in terms of image similarity, regional brain changes accuracy, and overall atrophy measured by the BPF, and assess their reliability in capturing the subtle and variable changes associated with healthy aging. As the results indicate, predicting brain changes during aging in a healthy population is indeed feasible, thereby answering our first research question. For almost all metrics, the best DF-based method outperformed the best GAN-based method. This suggests that DF-based methods remain superior for predicting longitudinal changes, as supported by our literature review.

### 5.1. Best Methods Comparison

Both visual and metrics results revealed that the DL and GAN methods effectively captured the volumetric changes of the ventricles. Similarly, the DF method accurately predicted changes ($p$-value < 0.01) in brain structures known to undergo marked changes in aging, particularly the thalamus and cortex (Choi et al., 2022; Fujita et al., 2023; Raz et al., 2005), as can be seen in

Table 6: Volume Fraction - Best methods Results

| Metric | lat. vent. ↑ | hippo. | thala. | cortex | ent. cortex | inf. temp. | sup. par. | mid. fron. | sup. fron. | precuneus | cuneus |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Baseline** | 1.83±0.8 | 0.69±0.1 | 1.07±0.1 | 37.5±0.8 | 0.26±0.0 | 2.25±0.1 | 1.57±0.2 | 1.77±0.1 | 3.85±0.2 | 1.41±0.1 | 0.68±0.1 |
| **Follow-up** | 2.50±1.2 | 0.68±0.1 | 1.04±0.1 | 36.7±0.9 | 0.26±0.0 | 2.14±0.1 | 1.54±0.2 | 1.68±0.1 | 3.70±0.2 | 1.44±0.1 | 0.70±0.1 |
| Best-DF | 2.47±1.0 | 0.70±0.1 | 1.03±0.1 | 37.3±0.8 | 0.29±0.0 | 2.24±0.1 | 1.54±0.1 | 1.74±0.1 | 3.75±0.2 | 1.40±0.1 | 0.68±0.1 |
| Best-GAN | 2.69±1.1 | 0.66±0.1 | 0.98±0.1 | 38.6±0.8 | 0.28±0.0 | 2.41±0.1 | 1.59±0.2 | 1.77±0.1 | 3.98±0.2 | 1.48±0.1 | 0.68±0.1 |

Volume Fraction (VF) of the selected tissues for the baseline and follow-up scans and the predictions of the best methods. ↑ indicates that the volume should increase with respect to the baseline; if no arrow is present, the volumes are expected to decrease. Underlined values indicate that the volume change is not possible with respect to the baseline.

Table 7: Dice Coefficient % - Best Methods Results

| Metric | lat. vent. | hippo. | thala. | cortex | ent. cortex | inf. temp. | sup. par. | mid. fron. | sup. fron. | precuneus | cuneus |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **LB** | 82.6±5.4 | 90.9±2.7 | 89.5±3.5 | 79.3±2.6 | 79.9±4.6 | 79.2±2.3 | 73.2±5.4 | 70.2±6.6 | 75.4±4.9 | 80.5±2.9 | 77.3±3.5 |
| Best-DF | **91.3±3.2** | ***91.1±2.3** | **93.4±1.9** | **83.0±2.5** | ***81.3±3.9** | **82.6±2.3** | **75.9±5.4** | **78.4±4.3** | **82.4±7.0** | **81.6±2.8** | ***78.0±3.5** |
| Best-GAN | 89.7±3.8 | 89.5±2.6 | 91.8±1.6 | 80.1±2.2 | 78.3±3.8 | 79.9±2.5 | 71.4±5.2 | 76.5±4.5 | 80.3±6.7 | *78.5±2.9 | *72.4±3.8 |

Dice coefficients of the selected tissues between the best methods and the follow-up scans. The initial Dice coefficient of the selected tissues between the baseline and follow-up scans is also shown as the Lower Bound (LB). The highest Dice coefficients between the best methods are indicated in **bold** (these values should also be higher than the LB). Values lower than the LB are underlined. * indicates $p$-values greater than 0.01. Note that higher Dice coefficients indicate better performance.

Table 8: Absolute Symmetrized Percent Volume Change - Best Methods Results

| Metric | lat. vent. | hippo. | thala. | cortex | ent. cortex | inf. temp. | sup. par. | mid. fron. | sup. fron. | precuneus | cuneus |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **UB** | 29.5±12 | 2.68±2.4 | 3.60±2.8 | 2.41±1.12 | 4.60±4.7 | 5.15±2.2 | 3.61±2.7 | 5.29±2.7 | 4.12±2.0 | 2.45±1.8 | 3.38±2.5 |
| Best-DF | **10.2±7.9** | 3.28±2.8 | **3.12±2.5** | **1.99±1.06** | 10.4±6.1 | **5.10±2.4** | ***2.99±2.1** | **3.87±2.4** | **1.79±1.7** | 2.81±2.0 | **3.23±2.5** |
| Best-GAN | 13.6±9.9 | 3.29±2.2 | *5.94±3.5 | 5.21±1.45 | 7.70±6.1 | ***12.0±3.1** | *4.59±3.5 | 5.44±2.7 | 7.37±3.0 | *3.54±2.6 | *3.83±2.9 |

Absolute Symmetrized Percent Volume Change (ASPVC) of the selected tissues between the best methods and the follow-up scans. The initial ASPVC of the selected tissues between the baseline and follow-up scans is also shown as the Upper Bound (UB). The lowest ASPVC values are indicated in **bold** (these values should also be lower than the UB). Values lower than the UB are underlined. * indicates $p$-values greater than 0.01. Note that Lower ASPVC values indicate better performance.

Tables 6, 7 and 8, demonstrating its capability to predict subtle changes in brain structures undergoing volume loss during aging. However, for other critical brain structures in aging, such as the hippocampus, entorhinal cortex, and precuneus, the DF method was unable to predict volume changes. This discrepancy may be due to the small size of these structures compared to the previous two, making them more challenging to predict. Additionally, there may be some segmentation errors as we observed unrealistic increases in volume in the actual follow-up scan in the precuneus and cuneus (see Table 6). In contrast, the GAN method did not show consistent predictions for any of these regions across the three metrics used, indicating a lack of sensitivity for brain structures other than the ventricles.

In the BPF analysis, the average results indicated a decrease in BPF in predictions made with the DF method, suggesting that brain atrophy was captured. However, in the GAN method, BPF tended to remain the same or even increase, which is unlikely in the aging brain of healthy individuals over a nine-year period (Fujita et al., 2023). The analysis showed that in the group with low BPF, the GANs results deviate much more from real predictions than the prediction by DL. This indicated that the method is less sensitive in individuals with accelerated brain aging. In contrast, predictions using the DF method showed that it was robust for all three BDF groups. These results were expected in the case of DF-based methods because, if an individual has low BPF (i.e., marked brain atrophy), the DF methods apply changes based on individuals with similarly low BPF, as these would be the most similar, thereby maintaining this trend in the prediction. The same principle applies to individuals with other BPF levels. The limitations of GANs may stem from the network's bias towards subjects with medium BPF fractions. Figure 7 shows that the means of the different groups in the GAN method are close to each other compared to the DF method or baseline/follow-up scans. A solution for this problem could be adding a hyperparameter to the network indicating that the individual has a high, medium, or low BPF at baseline, forcing the network to maintain appropriate BPF levels in predictions. This strategy has already been implemented in some studies predicting brain changes in patients with Alzheimer's disease (Ravi et al., 2019; Xia et al., 2021).

### 5.2. DF-Based Methods Analysis

A main finding in this family of methods was the validation of our hypothesis, that it is possible to use brain changes from known individuals to predict brain changes in others. The proof of our hypothesis is primarily shown in Table 3, but it can also be seen in Table 4 and Figure 4. Comparing the results of the best post-processed method with the upper bound indicates that registering with a DF obtained from individuals with similar structures yields results close to registering with the ground truth deformations.

As observed in Table 4, variations in results by chang-

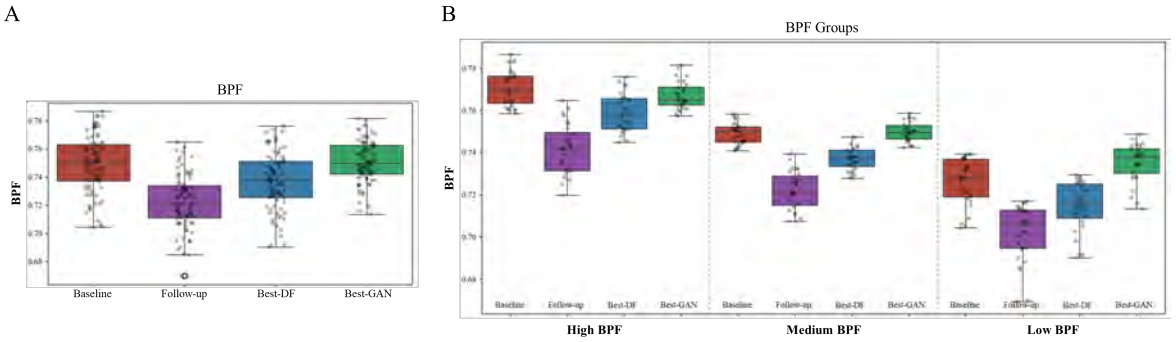



Figure 7: **Brain Parenchyma Fraction (BPF) - Best Methods Results.** (A) BPF of all individuals for the Baseline (Red), Follow-Up (Purple), Best DF-based Result (Blue), and Best GAN-based Result (Green). (B) BPF divided into groups by percentiles based on the BPF of the baseline: High BPF includes the 0-33 percentile (28 individuals), Medium BPF includes the 33-66 percentile (28 individuals), and Low BPF includes the 66-100 percentile (27 individuals).

ing hyperparameters were minimal, mainly affecting CSF and GM results slightly. However, the differences between local and non-local methods were much more pronounced, highlighting the potential of non-local methods to capture individual deformations and better adapt to the variability between individuals (Iglesias and Sabuncu, 2014), rather than calculating a global deformation average for the entire brain.

Despite the tissue-based method potentially being a more targeted approach for brain images, it did not yield better results than the patch-based method. This could be because the deformation field was obtained by non-overlapping tissues, leading to implausible deformations at the edges of each tissue due to abrupt changes that affects the inferred DF (Karacali and Davatzikos, 2006). A possible future solution could be to individually enlarge each tissue so they overlap and then calculate an average at their edges, which would avoid these abrupt deformations.

Another important point is that the B-spline method produced good visual results with low values for $n$. This suggests that performing a more exhaustive B-spline registration and slightly increasing $n$ could yield even better metrics. However, this would come with a significantly higher computational cost compared to other methods due to the extra registrations.

### 5.3. GAN-Based Methods Analysis

For this family of methods, one of the most significant finding was that the segmentation layer in the MPGAN+seg method outperformed the results of the MPGAN and all other GAN methods (see Table 5). Moreover, this was the only GAN method that did not worsen the lower bound. This demonstrated that guiding GANs with tissue losses is an effective approach for improving the accuracy of predictions in brain changes (Zhang et al., 2018).

An unexpected result was that training with 3D patches using MGAN yielded slightly better results than training with the full volume using MPGAN. More notably, the use of 2D slices with StyleGAN achieved superior results in both GM and the global image metric compared to the previous two methods. These results could be due to several reasons, but it is likely that one of the main factors was that training with smaller inputs allowed the creation of deeper networks that captured more image features and made more detailed predictions (Brown et al., 2020)

As seen in the results provided by the StyleGAN network, this network tried to preserve the individuals' identity, but there were still some notable changes in the overall brain shape that do not usually happen in brain aging of healthy individuals (see Figure 5). These problems were not found in the other GAN methos that used baseline-to-follow-up training with longitudinal images, allowing better maintenance of the global structure and the individual's identity (Peng et al. (2021), Huang et al. (2022))

### 5.4. Limitations of the Best Method

Despite the promising results by the best DF method, it still had some limitations.

First, the volume changes are restricted to possible variations within the population, making it impossible to capture individuals with changes outside this range. This limits the ability to observe abrupt changes, as most individuals in our population exhibit smaller changes.

Another limitation is that the dataset deformations have a specific resolution of 193x229x193, making it impossible to apply this method to new images with different dimensions without rescaling, which can lead to loss of detail. This issue can potentially be addressed by creating multi-resolution deformation datasets or by using deep learning techniques to resize the images, thereby reducing the loss of information (Umirzakova et al., 2023).

Finally, as mentioned initially, there was an inability

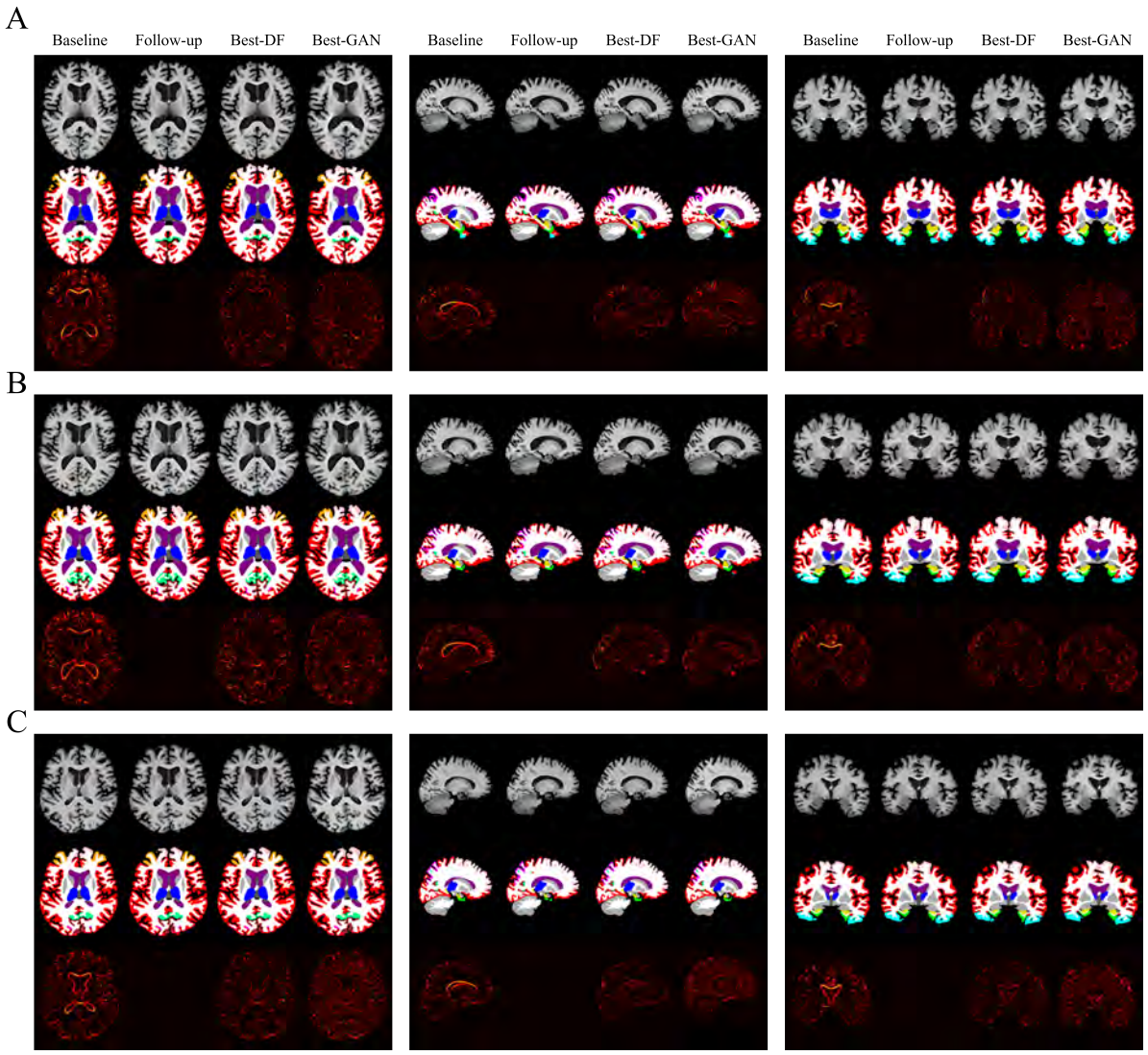



Figure 8: **Best Methods Predictions.** Baseline, Follow-Up, Best DF-based Prediction, and Best GAN-Based Prediction for three individuals. The axial plane is shown on the left, the sagittal plane in the center, and the coronal plane on the right. Each plane contains the prediction, the segmentation of the selected tissues (see Figure 6), and the difference with respect to the Follow-Up. (A) Individual with low BPF, (B) Individual with mean BPF, (C) Individual with high BPF.

to accurately predict changes in small brain regions such as the hippocampus, entorhinal cortex. This is a significant drawback, as these regions are crucial for the in-depth study of structural brain changes in aging (Fujita et al., 2023).

These three are the main limitations, although we know there may be others since this method has not been tested with images from other datasets.

### 5.5. Challenges of the Project

One of the main challenges of this project was attempting to predict the evolution of structural brain change with only two scans, assuming that the baseline scan had enough information in it to predict the follow-up. However, despite demonstrating that similar participants experienced similar brain aging, there were still specific changes in the brain of each participant that could only be calculated by having more time points between the baseline and the follow-up scans to measure the magnitude of changes for individuals in each brain region.

Another challenge was that the time between scans was quite long (around nine years). This causes much more variability between participants, as brain deformation is heavily affected by each individual's sociodemographic, health, genetics and lifestyle, and over nine years, many changes can occur (Mulugeta et al., 2022).

Another major challenge, was that most brain changes were quite subtle for most individuals. This led to very similar baseline and follow-up scans, making the visual evaluation of brain volume changes difficult.

### 5.6. Future Work

Future research could explore the integration of both strategies by introducing DF priors into GANs to guide volumetric changes. Additionally, incorporating diverse medical data from electronic health records or blood tests could further enhance the accuracy of these methods.

Enhancing GANs with segmentations that include a broader range of tissues, particularly those exhibiting significant changes during aging, could yield improved results. This could be accomplished by integrating an additional tissue segmentation network (Yu et al., 2022) and incorporating a loss function based on these tissues. However, this approach would necessitate substantially higher computational resources and result in slower training times.

Moreover, during this master's thesis, in collaboration with the computer science department, we experimented with a 2D diffusion model using autoencoders. The results were comparable to those obtained with the MGAN method but demonstrated greater stability during training. This suggests that future work focused on diffusion models holds significant promise.

## 6. Conclusions

This study investigated the prediction of structural brain changes in healthy adults over a nine-year period using 3D T1-weighted MRI images, comparing DF-based and GAN-based methods.

DF-based methods, based on the hypothesis that brain changes in some individuals can be used to predict changes in others individual from the same population, utilized multi-atlas techniques to combine volumetric changes from a subset of the population. Regional patch-based methods were the most effective.

We implemented four GAN methods based on recent work predicting brain structure changes in infants and patients with Alzheimer's disease, adapting them to our research questions. These methods aimed to train GANs to learn aging-related brain changes. However, most GAN methods were inaccurate in their predictions, with the exception of one model to which we added segmentation constraints.

Comparing the best methods from each family, DF-based methods outperformed GAN-based methods in nearly all metrics, capturing subtle changes in the thalamus and cortex. GAN methods predicted ventricular changes but lacked sensitivity for other structures. DF-based methods struggled with small regions like the hippocampus. DF-based methods were robust in predicting brain atrophy across varying BPF, while GAN methods were less accurate, especially for low BPF individuals.

This study provides a foundation for future research in brain change prediction, highlighting the effectiveness of DF-based methods and suggesting improvements for GAN methods. Future work could explore combining DF and GAN approaches, incorporating additional medical data, guiding GANs with more comprehensive segmentations, and exploring diffusion models.

## Acknowledgments

I would like to express my gratitude to Professors Xavier Llado and Arnau Oliver for their support during this project and their valuable advice on implementing the methods. I also wish to thank Professor Gabriel Kiss for allowing me to use the High Performance Computing cluster IDUN for deep learning training and for providing guidance on this family of methods. Additionally, I am grateful to Karl Hofseth, a fellow master's student, for testing diffusion models as an alternative approach. Finally, I would also like to extend my thanks to Dr. Live Eikenes for granting me access to the electronic materials needed to use FreeSurfer.

## References

Antipov, G., Baccouche, M., Dugelay, J.L., 2017. Face aging with conditional generative adversarial networks arXiv:1702.01983.

Arya, A., Verma, S., Chakarabarti, P., Chakrabarti, T., Elngar, A., Kamali, A.M., Nami, M., 2023. A systematic review on machine learning and deep learning techniques in the effective diagnosis of alzheimer's disease. Brain Informatics 10. doi:10.1186/s40708.023.00195.7.

Bandettini, P., 2012. Twenty years of functional mri: The science and the stories. NeuroImage 62, 575–88. doi:10.1016/j.neuroimage.2012.04.026.

Banerjee, S., Mittal, G., Joshi, A., Hegde, C., Memon, N., 2023. Identity-preserving aging of face images via latent diffusion models arXiv:2307.08585.

Bernal, J., Valverde, S., Kushibar, K., Cabezas, M., Oliver, A., Lladó, X., Alzheimer's Disease Neuroimaging Initiative, 2021. Generating longitudinal atrophy evaluation datasets on brain magnetic resonance images using convolutional neural networks and segmentation priors. Neuroinformatics 19, 477–492. doi:10.1007/s12021-020-09499-z.

Bethlehem, R.A., Seidlitz, J., White, S., Vogel, J., Anderson, K., Adamson, C., Adler-Wagstyl, S., Alexopoulos, G., Anagnostou, E., Areces Gonzalez, A., Astle, D., Auyeung, B., Ayub, M., Ball, G., Baron-Cohen, S., Beare, R., Bedford, S., Benegal, V., Beyer, F., Alexander-Bloch, A., 2021. Brain charts for the human lifespan doi:10.1101/2021.06.08.447489.

Brezova, V., Moen, K.G., Skandsen, T., et al., 2014. Prospective longitudinal mri study of brain volumes and diffusion changes during the first year after moderate to severe traumatic brain injury. NeuroImage: Clinical 5, 128–140. doi:10.1016/j.nicl.2014.03.012. published 2014 Mar 28.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language models are few-shot learners. arXiv:2005.14165.

Camara, O., Schweiger, M., Scahill, R., Crum, W., Sneller, B., Schnabel, J., Ridgway, G., Cash, D., Hill, D., Fox, N., 2006. Phenomenological model of diffuse global and regional atrophy using finite-element methods. IEEE Transactions on Medical Imaging 25, 1417–30. doi:10.1109/TMI.2006.880588.

Caruana, E., Roman, M., Hernández-Sánchez, J., Solli, P., 2015. Longitudinal studies. Journal of Thoracic Disease 7, E537–40. doi:10.3978/j.issn.2072-1439.2015.10.63.

Chen, X., Lathuilière, S., 2023. Face aging via diffusion-based editing arXiv:2309.11321.

Chen, Y., Almarzouqi, S.J., Morgan, M.L., Lee, A.G., 2018. T1-weighted image, in: Schmidt-Erfurth, U., Kohnen, T. (Eds.), Encyclopedia of Ophthalmology. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1747–1750. doi:10.1007/978.3.540.69000.9.1228.

Choi, E.Y., Tian, L., Su, J.H., Radovan, M.T., Tourdias, T., Tran, T.T., Trelle, A.N., Mormino, E., Wagner, A.D., Rutt, B.K., 2022. Thalamic nuclei atrophy at high and heterogenous rates during cognitively unimpaired human aging. NeuroImage 262, 119584. URL: https://www.sciencedirect.com/science/article/pii/S1053811922006991, doi:https://doi.org/10.1016/j.neuroimage.2022.119584.

Choi, Y., Uh, Y., Yoo, J., Ha, J.W., 2020. Stargan v2: Diverse image synthesis for multiple domains arXiv:1912.01865.

Coll, L., Pareto, D., Carbonell-Mirabent, P., Álvaro Cobo-Calvo, Arrambide, G., Ángela Vidal-Jordana, Comabella, M., Castilló, J., Rodríguez-Acevedo, B., Zabalza, A., Galán, I., Midaglia, L., Nos, C., Salerno, A., Auger, C., Alberich, M., Río, J., Sastre-Garriga, J., Oliver, A., Montalban, X., Àlex Rovira, Tintoré, M., Lladó, X., Tur, C., 2023. Deciphering multiple sclerosis disability with deep learning attention maps on clinical mri. NeuroImage: Clinical 38, 103376. doi:10.1016/j.nicl.2023.103376.

Crum, W., Hartkens, T., Hill, D., 2004. Non-rigid image registration: Theory and practice. The British Journal of Radiology 77 Spec No 2, S140–53. doi:10.1259/bjr/25329214.

Da Silva, M., Garcia, K., Sudre, C.H., Bass, C., Cardoso, M.J., Robinson, E., 2020. Biomechanical modelling of brain atrophy through deep learning arXiv:2012.07596.

Da Silva, M., Sudre, C.H., Garcia, K., Bass, C., Cardoso, M.J., Robinson, E.C., 2021. Distinguishing healthy ageing from dementia: a biomechanical simulation of brain atrophy using deep networks arXiv:2108.08214.

Duan, Y., Lin, Y., Rosen, D., Du, J., He, L., Wang, Y., 2020. Identifying morphological patterns of hippocampal atrophy in patients with mesial temporal lobe epilepsy and alzheimer disease. Frontiers in Neurology 11. doi:10.3389/fneur.2020.00021.

Fischl, B., 2012. Freesurfer. NeuroImage 62, 774–781. doi:10.1016/j.neuroimage.2012.01.021.

Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., Collins, D.L., 2011. Unbiased average age-appropriate atlases for pediatric studies. NeuroImage 54, 313–327. doi:10.1016/j.neuroimage.2010.07.033.

Fonov, V.S., Evans, A.C., McKinstry, R.C., Almli, C.R., Collins, D.L., 2009. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. NeuroImage 47, S102. doi:10.1016/S1053.8119(09)70884.5. organization for Human Brain Mapping 2009 Annual Meeting.

Fujita, S., Mori, S., Onda, K., Hanaoka, S., Nomura, Y., Nakao, T., Yoshikawa, T., Takao, H., Hayashi, N., Abe, O., 2023. Characterization of brain volume changes in aging individuals with normal cognition using serial magnetic resonance imaging. JAMA network open 6, e2318153. doi:10.1001/jamanetworkopen.2023.18153.

Gadewar, S., Zhu, A., Somu, S., Ramesh, A., Ba Gari, I., Thomopoulos, S., Thompson, P., Nir, T., Jahanshad, N., 2023. Normative aging for an individual's full brain mri using style gans to detect localized neurodegeneration, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2023, pp. 387–395. doi:10.1007/978.3.031.45676.3.39.

Ge, Y., Grossman, R.I., Babb, J.S., Rabin, M.L., Mannon, L.J., Kolson, D.L., 2002. Age-related total gray matter and white matter changes in normal adult brain. part i: Volumetric mr imaging analysis. American Journal of Neuroradiology 23, 1327–1333.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial networks arXiv:1406.2661.

Habes, M., Janowitz, D., Erus, G., Toledo, J.B., Resnick, S.M., Doshi, J., Van der Auwera, S., Wittfeld, K., Hegenscheid, K., Hosten, N., Biffar, R., Homuth, G., Völzke, H., Grabe, H.J., Hoffmann, W., Davatzikos, C., 2016. Advanced brain aging: relationship with epidemiologic and genetic risk factors, and overlap with alzheimer disease atrophy patterns. Translational Psychiatry 6, e775. doi:10.1038/tp.2016.39.

Hansen, T., Brezova, V., Eikenes, L., Håberg, A., Vangberg, T., 2015. How does the accuracy of intracranial volume measurements affect normalized brain volumes? sample size estimates based on 966 subjects from the hunt mri cohort. AJNR. American journal of neuroradiology 36. doi:10.3174/ajnr.A4299.

Hedman, A.M., van Haren, N.E., Schnack, H.G., Kahn, R.S., Hulshoff Pol, H.E., 2012. Human brain changes across the life span: A review of 56 longitudinal magnetic resonance imaging studies. Human Brain Mapping 33, 1987–2002. doi:https://doi.org/10.1002/hbm.21334.

Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., Reuter, M., 2020. Fastsurfer - a fast and accurate deep learning based neuroimaging pipeline. NeuroImage 219, 117012. doi:10.1016/j.neuroimage.2020.117012.

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.

Huang, X., Belongie, S., 2017. Arbitrary style transfer in real-time with adaptive instance normalization. arXiv:1703.06868.

Huang, Y., Ahmad, S., Han, L., Wang, S., Wu, Z., Lin, W., Li, G., Wang, L., Yap, P.T., 2022. Longitudinal prediction of postnatal brain magnetic resonance images via a metamorphic generative adversarial network arXiv:2208.04825.

Håberg, A.K., Hammer, T.A., Kvistad, K.A., et al., 2016. Incidental intracranial findings and their clinical impact; the hunt mri study in a general population of 1006 participants between 50-66 years. PLoS One 11, e0151080. doi:10.1371/journal.pone.0151080. published 2016 Mar 7.

Iglesias, J., Sabuncu, M., 2014. Multi-atlas segmentation of biomedical images: A survey. Medical Image Analysis 24. doi:10.1016/j.media.2015.06.012.

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, eadd3607. doi:10.1126/sciadv.add3607.

Iglesias, J.E., Billot, B., Balbastre, Y., Tabari, A., Conklin, J., González, R.G., Alexander, D.C., Golland, P., Edlow, B.L., Fischl, B., 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. doi:10.1016/j.neuroimage.2021.118206.

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2018a. Image-to-image translation with conditional adversarial networks arXiv:1611.07004.

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2018b. Image-to-image translation with conditional adversarial networks. arXiv:1611.07004.

Karacali, B., Davatzikos, C., 2006. Simulation of tissue atrophy using a topology preserving transformation model. IEEE Transactions on Medical Imaging 25, 649–652. doi:10.1109/TMI.2006.873221.

Karnewar, A., Wang, O., 2020. Msg-gan: Multi-scale gradients for generative adversarial networks. arXiv:1903.06048.

Kaye, J.A., DeCarli, C., Luxenberg, J.S., Rapoport, S.I., 1992. The significance of age-related enlargement of the cerebral ventricles in healthy men and women measured by quantitative computed x-ray tomography. Journal of the American Geriatrics Society 40, 225–231. doi:10.1111/j.1532-5415.1992.tb02073.x.

Khanal, B., Ayache, N., Pennec, X., 2017. Simulating longitudinal brain mris with known volume changes and realistic variations in image intensity. Frontiers in Neuroscience 11. doi:10.3389/fnins.2017.00132.

Khanal, B., Lorenzi, M., Ayache, N., Pennec, X., 2016. A biophys-

ical model of brain deformation to simulate and analyse longitudinal mris of patients with alzheimer's disease. NeuroImage 134. doi:10.1016/j.neuroimage.2016.03.061.

Kingma, D.P., Ba, J., 2017. Adam: A method for stochastic optimization. arXiv:1412.6980.

Klein, S., Staring, M., Murphy, K., Viergever, M., Pluim, J., 2009. Elastix: A toolbox for intensity-based medical image registration. IEEE transactions on medical imaging 29, 196–205. doi:10.1109/TMI.2009.2035616.

Lee, J., Mustafaev, T., Nishikawa, R., 2023. Impact of gan artifacts for simulating mammograms on identifying mammographically occult cancer. Journal of Medical Imaging 10. doi:10.1117/1.JMI.10.5.054503.

Li, S., Lei, H., Zhou, F., Gardezi, J., Lei, B., 2019. Longitudinal and multi-modal data learning for parkinson's disease diagnosis via stacked sparse auto-encoder, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 384–387. doi:10.1109/ISBI.2019.8759385.

Modat, M., Simpson, I., Cardoso, M.J., Cash, D., Toussaint, N., Fox, N., Ourselin, S., 2014. Simulating neurodegeneration through longitudinal population analysis of structural and diffusion weighted mri data, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014, pp. 57–64. doi:10.1007/978.3.319.10443.0.8.

Mulugeta, A., Navale, S.S., Lumsden, A.L., Llewellyn, D.J., Hyppönen, E., 2022. Healthy lifestyle, genetic risk and brain health: A gene-environment interaction study in the uk biobank. Nutrients 14. doi:10.3390/nu14193907.

Peng, L., Lin, L., Lin, Y., Chen, Y.w., Mo, Z., Vlasova, R.M., Kim, S.H., Evans, A.C., Dager, S.R., Estes, A.M., McKinstry, R.C., Botteron, K.N., Gerig, G., Schultz, R.T., Hazlett, H.C., Piven, J., Burrows, C.A., Grzadzinski, R.L., Girault, J.B., Shen, M.D., Styner, M.A., 2021. Longitudinal prediction of infant mr images with multi-contrast perceptual adversarial learning. Frontiers in Neuroscience 15. doi:10.3389/fnins.2021.653213.

Pini, L., Pievani, M., Bocchetta, M., Altomare, D., Bosco, P., Cavedo, E., Galluzzi, S., Marizzoni, M., Frisoni, G.B., 2016. Brain atrophy in alzheimer's disease and aging. Ageing Research Reviews 30, 25–48. doi:10.1016/j.arr.2016.01.002. brain Imaging and Aging.

Pintzka, C.W., Hansen, T.I., Evensmoen, H.R., Håberg, A.K., 2015. Marked effects of intracranial volume correction methods on sex differences in neuroanatomical structures: a hunt mri study. Frontiers in Neuroscience 9. doi:10.3389/fnins.2015.00238.

Rachmadi, M., Valdés-Hernández, M., Makin, S., Wardlaw, J., Komura, T., 2019. Predicting the evolution of white matter hyperintensities in brain mri using generative adversarial networks and irregularity map, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2019, pp. 146–154. doi:10.1007/978.3.030.32248.9.17.

Ravi, D., Alexander, D.C., Oxtoby, N.P., 2019. Degenerative adversarial neuroimage nets: Generating images that mimic disease progression arXiv:1907.02787.

Raz, N., Lindenberger, U., Rodrigue, K.M., Kennedy, K.M., Head, D., Williamson, A., Dahle, C., Gerstorf, D., Acker, J.D., 2005. Regional brain changes in aging healthy adults: General trends, individual differences and modifiers. Cerebral Cortex 15, 1676–1689. doi:10.1093/cercor/bhi044.

Reuter, M., Rosas, H.D., Fischl, B., 2010. Highly accurate inverse consistent registration: A robust approach. NeuroImage 53, 1181–1196. doi:10.1016/j.neuroimage.2010.07.020.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, Cham, pp. 234–241.

Rudick, R.A., Fisher, E., Lee, J.C., Simon, J., Jacobs, L., 1999. Use of the brain parenchymal fraction to measure whole brain atrophy in relapsing-remitting ms. multiple sclerosis collaborative research group. Neurology 53, 1698–1704. doi:10.1212/wnl.53.8.1698.

Schulz, M., Mayer, C., Schlemm, E., Frey, B., Malherbe, C., Petersen, M., Gallinat, J., Kühn, S., Fiehler, J., Hanning, U., Twerenbold, R., Gerloff, C., Cheng, B., Thomalla, G., 2022. Association of age and structural brain changes with functional connectivity and executive function in a middle-aged to older population-based cohort. Frontiers in Aging Neuroscience 14. doi:10.3389/fnagi.2022.782738.

Sharma, S., Noblet, V., Rousseau, F., Heitz, F., Rumbach, L., Armspach, J.P., 2010. Evaluation of brain atrophy estimation algorithms using simulated ground-truth data. Medical Image Analysis 14, 373–89. doi:10.1016/j.media.2010.02.002.

Smith, A., Crum, W., Hill, D., Thacker, N., Bromiley, P., 2003. Biomechanical simulation of atrophy in mr images. Proceedings of SPIE 5032, 481–490. doi:10.1117/12.480412.

Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S., 2015. Deep unsupervised learning using nonequilibrium thermodynamics arXiv:1503.03585.

Umirzakova, S., Mardieva, S., Muksimova, S., Ahmad, S., Whangbo, T., 2023. Enhancing the super-resolution of medical images: Introducing the deep residual feature distillation channel attention network for optimized performance and efficiency. Bioengineering 10. URL: https://www.mdpi.com/2306-5354/10/11/1332, doi:10.3390/bioengineering10111332.

Vemuri, P., Murray, M.E., Jack, C.R., 2015. Chapter 10 - neuroimaging in dementias, in: Rosenberg, R.N., Pascual, J.M. (Eds.), Rosenberg's Molecular and Genetic Basis of Neurological and Psychiatric Disease (Fifth Edition). fifth edition ed.. Academic Press, Boston, pp. 107–118. doi:10.1016/B978.0.12.410529.4.00010.3.

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. doi:10.1109/TIP.2003.819861.

Xia, T., Chartsias, A., Wang, C., Tsaftaris, S.A., 2021. Learning to synthesise the ageing brain without longitudinal data. Medical Image Analysis 73, 102169. doi:10.1016/j.media.2021.102169.

Yu, X., Tang, Y., Zhou, Y., Gao, R., Yang, Q., Lee, H.H., Li, T., Bao, S., Huo, Y., Xu, Z., Lasko, T.A., Abramson, R.G., Landman, B.A., 2022. Characterizing renal structures with 3d block aggregate transformers. arXiv:2203.02430.

Zapaishchykova, A., Tak, D., Ye, Z., Liu, K.X., Likitlersuang, J., Vajapeyam, S., Chopra, R.B., Seidlitz, J., Bethlehem, R.A.I., Mak, R.H., Mueller, S., Haas-Kogan, D.A., Poussaint, T.Y., Aerts, H.J.W.L., Kann, B.H., 2024. Diffusion deep learning for brain age prediction and longitudinal tracking in children through adulthood. Imaging Neuroscience 2, 1–14. doi:10.1162/imag.a.00114.

Zhang, Z., Yang, L., Zheng, Y., 2018. Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9242–9251. doi:10.1109/CVPR.2018.00963.

Zhou, Z., Sodha, V., Siddiquee, M.M.R., Feng, R., Tajbakhsh, N., Gotway, M.B., Liang, J., 2019. Models genesis: Generic autodidactic models for 3d medical image analysis arXiv:1908.06912.

Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2020. Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv:1703.10593.

Åsvold, B., Langhammer, A., Rehn, T., Kjelvik, G., Grøntvedt, T., Sørgjerd, E., Fenstad, J., Heggland, J., Holmen, O., Stuifbergen, M., Aalberg Vikjord, S., Brumpton, B., Skjellegrind, H., Thingstad, P., Sund, E., Selbæk, G., Mork, P., Rangul, V., Hveem, K., Krokstad, S., 2022. Cohort profile update: The hunt study, norway. International Journal of Epidemiology 52. doi:10.1093/ije/dyac095.

# A. T1w MR Images Details

Table 9: MRI Sequence Parameters

| Dataset | Matrix size | NSA | TR (ms) | TE (ms) | Flip-angle | Slice thickness (mm) | Gap (mm) | Overlap (mm) | FOV (mm) |
|---|---|---|---|---|---|---|---|---|---|
| HUNT3 | 192x192 | 1 | 10.2 | 4.1 | 10° | 1.2 | 0 | 0 | 240 |
| HUNT4 | 256x192 | - | 7.7 | 3.092 | 8° | 1.0 | 0 | 0 | 256 |

Parameters of the MRI sequence for HUNT3 and HUNT4 dataset, including matrix size, number of signal averages (NSA), repetition time (TR), echo time (TE), flip-angle, slice thickness, gap, overlap, and field of view (FOV).

# B. Used Parameter for Non-Rigid Registration

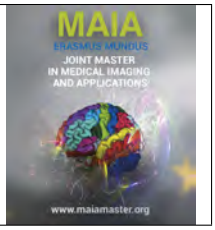
Table 10: Parameters Used in the B-Spline Transformation

| Parámetro | Valor |
|---|---|
| UseDirectionCosines | true |
| Registration | MultiMetricMultiResolutionRegistration |
| Interpolator | BSplineInterpolator |
| ResampleInterpolator | FinalBSplineInterpolator |
| Resampler | DefaultResampler |
| FixedImagePyramid | FixedRecursiveImagePyramid |
| MovingImagePyramid | MovingRecursiveImagePyramid |
| Optimizer | AdaptiveStochasticGradientDescent |
| Transform | BSplineTransform |
| Metric | AdvancedNormalizedCorrelation, TransformBendingEnergyPenalty |
| FinalGridSpacingInVoxels | 4 4 4 |
| NumberOfHistogramBins | 32 |
| Metric0Weight | 1.0 |
| Metric1Weight | 0.1 |
| NumberOfResolutions | 2 |
| ImagePyramidSchedule | 1 1 1 1 1 1 |
| MaximumNumberOfIterations | 1000 |
| MaximumStepLength | 0.117188 |
| NumberOfSpatialSamples | 2048 |
| ImageSampler | Random |
| BSplineInterpolationOrder | 1 |
| FinalBSplineInterpolationOrder | 3 |

Most important parameters used in the B-Spline registration to create the DF dataset used in the DF-based family.

Medical Imaging and Applications

Master Thesis, June 2024




# CMR-to-CTA Image Conversion using Diffusion Models for Transcatheter Aortic Valve Implantation Planning

Carmen Guadalupe Colin-Tenorio, Agnes Mayr, Christian Kremser, Markus Haltmeier, Enrique Almar-Munoz

*Medical University of Innsbruck, Austria*

**Abstract**

***Introduction:*** Transcatheter Aortic Valve Implantation (TAVI) has become the preferred method for treating severe aortic stenosis, especially in patients who are unsuitable for traditional surgery. Typically, preoperative imaging for TAVI involves contrast-enhanced Computed Tomography Angiography (CTA). However, for patients with contraindications to contrast agents, cardiac magnetic resonance imaging (CMR) is a viable alternative, albeit with its limitations in visualizing calcifications.

***Methods:*** This study explores the application of diffusion models to enhance CMR-to-CTA image conversion, facilitating comprehensive TAVI planning without needing contrast agents. We developed a pipeline incorporating Denoising Diffusion Probabilistic Models (DDPMs) and Score-Matching to synthesize CTA-equivalent images from CMR scans. This approach was evaluated using an in-house dataset of 39 paired CTA and CMR scans from the Tirol Kliniken (Innsbruck, Austria) database.

***Results:*** Our results show that the synthesized CTA images maintain high fidelity to their real counterparts, as validated by metrics such as the Structural Similarity Index Measure (SSIM) and Peak Signal-to-Noise Ratio (PSNR), with values above 0.80 and 22 respectively.

***Conclusion:*** This study highlights the potential of diffusion models in medical imaging, offering a promising solution for patients unable to receive contrast agents, thereby improving the safety and efficacy of TAVI procedures.

*Keywords:* Diffusion Models, Segmentation, Registration, CMR-CTA Conversion, Transcatheter Aortic Valve Intervention

## 1. Introduction

Aortic Stenosis (AS) refers to the narrowing of the aortic valve opening and can sometimes be referred to as a failing heart valve. This condition restricts blood flow from the left ventricle to the aorta, which may also impact pressure in the left atrium. While some individuals may have aortic stenosis due to a congenital heart defect known as a bicuspid aortic valve, this condition more commonly develops during aging as calcium or scarring accumulates, causing damage to the valve and restricting blood flow. This pathology represents the most prevalent valvular abnormality in the Western world, with severe AS affecting 3% of individuals aged 75 or older (Lindroos et al., 1993). Despite its prevalence, a substantial proportion of these patients are ineligible for aortic valve replacement (AVR) due to high surgical risk (Fanning et al., 2013).

Transcatheter aortic valve implantation (TAVI), a minimally invasive technique, has emerged as the gold standard for treating severe aortic stenosis in patients unsuitable for surgery or at high surgical risk (Lukas et al., 2022). It involves the insertion of a new valve through a catheter, which is typically entered through the femoral artery and guided to the heart. Once the catheter reaches the aortic valve, the new valve is positioned inside the diseased valve and expanded. This effectively displaces the old, narrowed valve and allows the new valve to take over the function of regulating blood flow from the heart to the aorta and the rest of the body.

The standard pre-imaging workup for TAVI planning includes Transthoracic and Transesophageal Echocar-

diography, alongside contrast-enhanced Computed Tomography Angiography (CTA), to precisely determine valve size and implantation route (Al-Najafi et al., 2016). Notably, up to 80% of patients undergoing TAVI suffer from chronic renal insufficiency

However, contrast agents are contraindicated in patients with acute or chronic kidney disease (CKD), with prevalence rates up to 41% for acute kidney injuries and up to 70% for CKD in TAVI patients (Jhaveri et al., 2017; Ram et al., 2017). Consequently, there is a critical need for contrast-free methods for TAVI planning.

TAVI applications of cardiovascular magnetic resonance (CMR) are emerging. CMR can provide the structural and functional imaging details required for TAVI procedure (Mahon and Mohiaddin, 2021). CMR is a viable alternative to CTA, offering comprehensive TAVI planning without the need for iodinated contrast media and radiation exposure. Despite challenges in delineating vascular calcifications with CMR, studies have shown that CMR-guided TAVI is comparable to CTA-guided TAVI in terms of implantation success (Mayr et al., 2018). For instance, Pamminger et al. (2020) demonstrated that unenhanced quiescent-interval single-shot MR angiography (QISS-MRA) combined with 3D "whole heart" CMR protocols can facilitate fully unenhanced TAVI guidance. However, QISS-MRA does not visualize calcified plaque burden, depicting only the vessel lumen and not the vessel wall.

An essential consideration is the degree of calcification. While CT may overestimate calcium in heavily calcified valves and arteries, CMR may underestimate it (Barbanti et al., 2013). Vascular calcifications produce very low signal intensity with standard CMR pulse sequences because of their low free water concentration and short T2*. This makes them challenging to visualize on CMR. The discrepancy could be due to the blooming artifact that artificially enlarges dense calcifications on CT images. Another possibility is that the surface regions of a calcification contain mobile water spins, which could generate detectable signal intensity and therefore decrease their apparent volume with PDIP-SOS CMR (Serhal et al., 2018).

Figure 1 shows an example of how the aortic valve appears in both modalities, with calcifications appearing white in the CT image and black in the CMR image. Thus, integrating CMR and CT is desirable for comprehensive TAVI planning, though using both modalities simultaneously remains an active research area.

Image synthesis across and within medical imaging modalities is an evolving field with broad applications in radiology. Its primary purpose is to streamline clinical workflows by bypassing or replacing an imaging procedure when the acquisition is infeasible due to constraints like contraindications to ionizing radiation. Recent advancements in deep learning have enabled the development of methods that can be generalized across

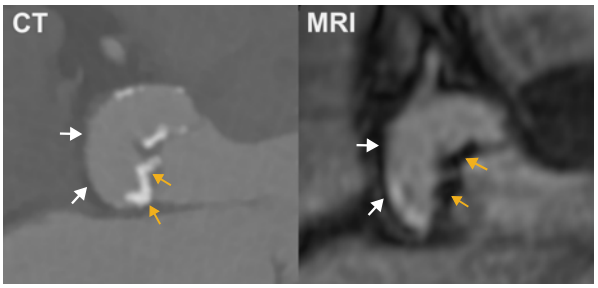



Figure 1: Comparison of Aortic Valve Visualization in CT and CMR Registered Images. In both modalities, calcifications are present (yellow arrows); they appear white in the CT image and black in the CMR image. White arrows show the border of the valve

different pairs of imaging modalities with minimal adjustments (Wang et al., 2021). Particularly, several deep learning-based cross-modality medical image synthesis studies have utilized convolutional neural networks (CNNs) and generative adversarial networks (GANs) (Lyu and Wang, 2022).

While GANs have been the state-of-the-art for synthetic image generation due to their high image quality, they suffer from instability during training and low diversity generation due to mode collapse (Kazerouni et al., 2023). Recently, denoising diffusion models have emerged in computer vision, demonstrating remarkable results in generative modeling, including applications in medical imaging. These models generate high-fidelity, realistic images and outperform GANs and variational autoencoders in multiple image generation tasks (Croitoru et al., 2023).

## 2. State of the Art

### *2.1. Diffusion Models*

Diffusion models represent a cutting-edge class of generative models that have proven highly effective in learning complex data distributions (Croitoru et al., 2023). Unlike other generative models such as GANs and variational autoencoders, which are challenging to train and interpret and often do not produce satisfactory image quality, diffusion models are analytically principled and straightforward to train. They exhibit impressive generative capabilities, producing high-detail and diverse examples. Studies increasingly show that diffusion models outperform GANs and variational autoencoders in various image generation tasks (Croitoru et al., 2023). Denoising Diffusion Probabilistic Models (DDPMs) are a class of generative models designed to produce high-quality synthetic images. Introduced by Ho et al. (2020), these models operate by iteratively adding and then removing noise from an image. During training, DDPMs learn to predict the noise added to images through a gradual, step-by-step process. At inference, the model starts with a noisy image and sys-

tematically denoises it to generate a clear, realistic image.

### *2.2. Diffusion Models in Medical Imaging*

Several significant works have explored the use of diffusion models for generating synthetic medical images. For instance, Dorjsembe et al. (2022) applied the original pipeline of DDPMs for generating high-quality CMR images of brain tumors. Similarly, Txurio et al. (2023) applied diffusion models for realistic CT image generation, while Pan et al. (2023b) demonstrated the generation of high-quality 2D medical images across different imaging modalities using a transformer-based DDPM. Their framework leverages a Swin-transformer-based network for the denoising process, allowing the creation of realistic and diverse synthetic images from datasets, including chest X-rays, heart CMR, pelvic CT, and abdomen CT. Luo and Hu (2024) proposed measurement-guided diffusion models for high-quality medical image synthesis, ensuring generation stability and improving learning ability. Bradbury et al. (2024) introduced a novel technique for generating related, synthetic PET-CT-Segmentation scans. The method employs linked DDPMs, enabling paired diffusion to enhance imaging consistency and segmentation accuracy. Ricardo et al. (2023) introduced MAM-E, a pipeline of generative models for high-quality mammographic image synthesis, capable of generating images based on a text prompt description and also capable of generating lesions using stable diffusion.

### *2.3. Image Synthesis in Multimodality*

In the realm of multimodality image synthesis, Pan et al. (2023a) proposed synthetic CT generation from CMR using a 3D transformer-based denoising diffusion model. This method addresses the challenge of aligning the anatomical structures captured by different imaging modalities, ensuring that the synthetic images are accurate and clinically useful. Additionally, Graf et al. (2023) demonstrated that DDPMs can be used for CMR to CT image translation, significantly enhancing automated spinal segmentation. This approach involves aligning spinal CMR and CT images through rigid landmark registration, training image-to-image translation models, and subsequently generating synthetic CT images from CMR.

Furthermore, Lyu and Wang (2022) proposed a novel approach for CT-CMR conversion specifically for pelvis images. This study utilized a DDPM conditioned and score-matching framework to generate realistic CT images from CMR, addressing the challenges of anatomical structure preservation and realistic texture synthesis.

### *2.4. Diffusion Models in Cardiac Imaging*

Stojanovski et al. (2023) proposed using DDPM for the generation of synthetic ultrasound images guided by cardiac diffusion models, aimed at improving real image segmentation. Their work demonstrates the potential of diffusion models in creating synthetic cardiac ultrasound images that can aid in various clinical applications. Hantao et al. (2024) introduced a technique for synthesizing myocardial pathology on cardiac CMR scans using lesion-focus diffusion models. This method accurately models cardiac lesions, enhancing diagnostic capabilities and providing realistic pathological scenarios for medical training and evaluation.

While extensive research has been conducted on diffusion models for various medical imaging applications, further exploration in cardiac imaging is needed. Studies focusing on this area could provide valuable insights and potentially enhance the current methodologies for cardiac image synthesis and analysis, paving the way for improved TAVI planning and other cardiac procedures.

### *2.5. Project Description*

The objective of this project is to explore the use of diffusion models, specifically, Denoising Diffusion Probabilistic Models (DDPM) and score matching, conditioned on medical imaging data. We aim to develop an innovative approach for converting CMR images into CT-equivalent images. This conversion process will enable TAVI surgeons to leverage the benefits of both imaging modalities without the need for risky contrast agents. The primary goal is to use CMR as the input and generate the corresponding CT image. To the best of our knowledge, this is the first MR-CT Image conversion method for cardiac images, an application specially difficult due to the breathing and heart-beating movements. The contributions of this work are:

- *Providing a Solution for Patients Unable to Receive Contrast Agents*: We explored a potential solution for TAVI patients who cannot receive contrast agents and only have CMR available for their procedure.
- *Dataset Construction*: We construct a dataset comprising co-registered CMR and CT image pairs with aorta segmentations.
- *Application of Diffusion Models*: We apply diffusion models to generate CT scans from CMR images and compare the generated images with the real CT images to evaluate the effectiveness of our approach.

## 3. Material and methods

This section describes the complete methodology employed in this work, covering dataset acquisition, preprocessing, aorta segmentation, registration methods, image generation, and quantitative criteria for evaluation. Figure 2 provides a comprehensive overview of the pipeline developed.

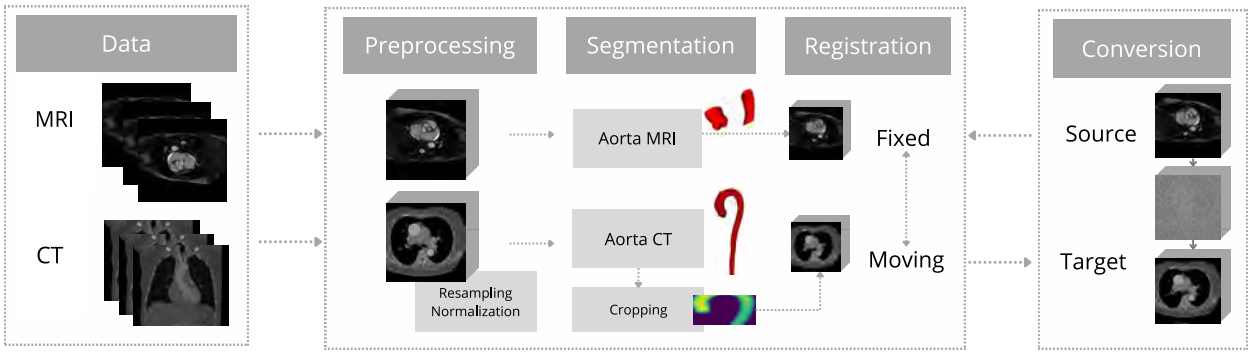



Figure 2: Pipeline description

### 3.1. Dataset

#### 3.1.1. Dataset Acquisition

For this study, we acquired a total of 147 CT scans and their corresponding CMR scans directly in DICOM format from the in-house database of the Tirol Kliniken (Innsbruck, Austria). The scans were taken between 2015 and 2022, and both modalities were captured in the diastole phase. CT protocols included scans of the body trunk and heart.

#### 3.1.2. Dataset Selection

To build the dataset, stringent criteria were created to ensure the acquisition of good-quality and comparable multimodal images, which is essential for robust analysis.

**Inclusion Criteria:**

- CT and CMR volumes cover all the heart and the aortic valve region.
- To minimize anatomical variations, the time gap between the CT and CMR scans should be inferior than one year.
- Keep only CT taken with an iodinated contrast media to ensure clear visualization of the aortic valve leaflets, as without contrast, only calcifications are visible. Our goal is to replicate them computationally.
- Exclusion of CT scans with significant artifacts or those obtained post-TAVI procedure.

Figure 3 illustrates the patient selection process, where 39 patients with both CT and CMR scans met the inclusion criteria.

#### 3.1.3. Data preprocessing

We used linear interpolation to resample all the scans to a 1mm x 1mm pixel size.

In Figure 4, the image orientation and volume coverage for both CMR and CT are illustrated. The yellow

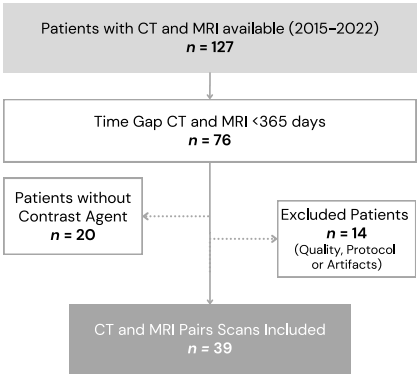



Figure 3: Data flow of the patient selection

lines in the CMR (left) shows the range and orientation focusing specifically on the aortic valve, where it can be seen the orientation is perpendicular to the aortic valve, and the range includes the complete aortic valve, but not the whole aorta. The CT image on the right encompasses the whole body trunk, in this case the yellow lines are including the whole aorta, descendent and ascendant, and the orientation is perpendicular to the large axis of the body.

In the CMR acquisition a parameter to choose from is the flip angle, which is chosen according to the patient anatomy and it is computed perpendicularly to the ascendant aorta. In our images, flip angles were from 22 to 70. The CTA protocol is without a flip angle.

Subsequently, we preprocessed the images to ensure that they were equally oriented in the RAS (Right-Anterior-Superior) configuration. The next step was to train the aorta segmentation models for each image modality.

### 3.2. Aorta Segmentation

Using aorta segmentation masks to register CTA-CMR is effective because the aorta is a relatively rigid and stable structure, maintaining consistent shape and position, especially if both modalities are taken in the same cardiac phase. As the aorta presents an ascendant

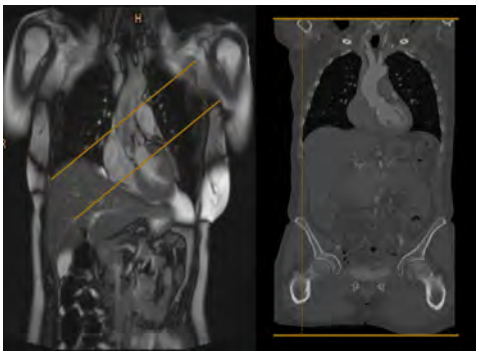


Figure 4: Typical image orientation and volume coverage of whole heart 3D MRA (left) and whole aorta CT (right). The orange lines indicate the range and orientation of acquired CMR images. The range is focused specifically on the aortic valve.

and descendant part, their relative position simplifies the alignment process, reducing errors. The ease of segmenting the aorta in both modalities makes it an ideal reference for accurate and reliable image registration.

Prior to training the model, two expert radiologists were tasked to manually annotate 40% of the scans for training and 20% for validation, 16 and 8 volumes respectively. The remaining volumes were kept for inference once the model was properly trained.

#### *3.2.1. CMR Aorta Segmentation*

The nnU-Net framework, was employed for automatic aorta segmentation due to its ability to self-configure preprocessing, network architecture, and post-processing pipelines for medical image segmentation (Isensee et al., 2021). This framework adapts its parameters based on the dataset provided, optimizing segmentation tasks without extensive manual adjustments. This study utilized two nnU-Net models: 2D Unet and 3D Unet.

No modifications were made in setting the nnU-Net hyperparameters and data augmentation strategy. A 5-fold cross-validation strategy was applied throughout the training to fully utilize the patient data.

#### *3.2.2. CTA Aorta Segmentation*

For the CTA-based Aorta Segmentation we are using the same nnUNet architecture, but we are comparing two training configurations:

- **Self-trained nnUNet network:** We use the same training strategy as in Sec. 3.2.1, but with CTA images.

- **Pre-trained nnUNet network:** We used the pre-trained TotalSegmentator framework, Wasserthal et al. (2022). It is a nnU-Net model pretrained on 1204 CT scans and segments 104 structures.

#### *3.2.3. Cropping*

The dimensions of the CMR slices ranged from (512, 384, 72) to (512, 384, 112), whereas the CT slices varied from (512, 512, 201) to (512, 512, 952). Due to these distinct differences in slice ranges, with CMR predominantly capturing the aortic valve, a standardized cropping method was necessary. We employed the aorta segmentation obtained from the previous step for cropping the long-acquired volumes.

To achieve this, the aorta segmentation was projected from 3D to 2D to identify the region with the highest density of information. An example of this projection is illustrated in Figure 5, showing the highest density points marked by a red 'X'. Using the coordinates derived from this projection, we cropped the CT images, retaining 30 slices above and below the identified point. This approach ensured that the aorta region in the CT images was comparable to the aortic information captured in the CMR.

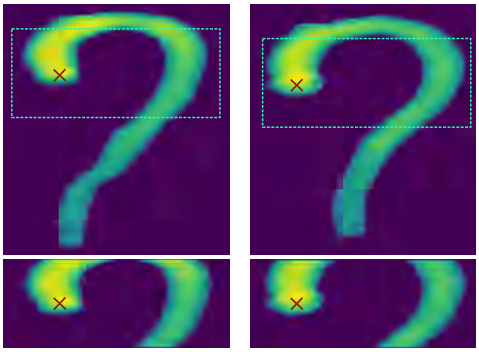


Figure 5: Example of the 2D projection of the aorta segmentation, before (above) and after cropping (below), where the highest density points is marked by the red 'X'

### *3.3. Registration*

The registration methods can further be classified as rigid, affine, or deformable, depending on the nature of the transformations allowed.

In our work, we focus on multimodal intra-patient registration based on 3D volume images between CT and CMR. The goal of this step is to register the CT slices with respect to their corresponding CMR images, ensuring that the anatomical structures align correctly across the two modalities.

#### *3.3.1. Deep Learning-based registration*

For precise image registration, we employed two state-of-the-art deep learning frameworks: VoxelMorph, (Balakrishnan et al., 2019), and TransMorph, (Chen et al., 2022). These models are specifically designed to handle complex deformable image registration tasks.

VoxelMorph performs the registration process by learning a mapping from an input image pair to a deformation field that aligns the images. This mapping is parameterized using a convolutional neural network (CNN). TransMorph, in contrast, utilizes a transformer-based architecture to enhance registration performance by capturing both global and local contexts. This model learns complex spatial relationships within the images, allowing for a more detailed and accurate mapping of the deformation field. For our experiments, we used the Tiny version of TransMorph. We experimented with various loss functions and included the aorta segmentation masks during training. In our study, both VoxelMorph and TransMorph were applied to register CT (fixed) and CMR (moving) images along with their corresponding segmentation masks. This configuration was performed because in our images the CTA covers a larger area than the CMR, in the axial view, so it is better for the CMR to be the moving image to reduce the number of empty pixels after registration.

The loss function for TransMorph and VoxelMorph consists of three components: similarity, regularization, and Dice loss. The similarity term measures the similarity between the deformed moving image and the fixed image, the regularization term ensures the smoothness of the deformation field, and the Dice loss term assesses the overlap between the predicted and true segmentation masks. In each component, there is a coefficient from 0 to 1 to choose the weight of each component at the final loss.

The loss function is defined as follows:

$$\mathcal{L} = \mathcal{L}_{\text{similarity}} + \mathcal{L}_{\text{regularizer}} + \mathcal{L}_{\text{Dice}}$$

We experimented with two similarity metrics specifically for multimodal registration: Mutual Information (MI) and Modality Independent Neighborhood Descriptor (MIND), to capture image similarity. The mutual information $I(A; B)$ between two images $A$ and $B$ is given by:

$$I(A; B) = \sum_{a \in A} \sum_{b \in B} p(a, b) \log \left( \frac{p(a, b)}{p(a)p(b)} \right)$$

Where:

- $p(a, b)$ is the joint probability distribution of the pixel intensities in images $A$ and $B$.
- $p(a)$ and $p(b)$ are the marginal probability distributions of the pixel intensities in images $A$ and $B$, respectively.

The Modality Independent Neighbourhood Descriptor (MIND) for a voxel $\mathbf{p}$ in image $I$ is defined as:

$$\text{MIND}_{\mathbf{p}}(\mathbf{q}) = \exp\left( -\frac{\text{D}_{\mathbf{p}}(\mathbf{q})}{V_{\mathbf{p}}} \right)$$

Where:

- $\mathbf{q}$ is a neighboring voxel of $\mathbf{p}$.
- $\text{D}_{\mathbf{p}}(\mathbf{q})$ is the squared Euclidean distance between the patches centered at $\mathbf{p}$ and $\mathbf{q}$
- $V_{\mathbf{p}}$ is the variance of the distances $\text{D}_{\mathbf{p}}(\mathbf{q})$ within the local neighborhood $\mathcal{N}$.

The CMR registered images needed to maintain the anatomical structures, since they will serve for planning the TAVI procedure, and to choose the correct device by measuring those anatomical structures. For this reason, other types of registration were explored.

### *3.3.2. Traditional Methods for registration*

Elastix is an open-source software package widely used for medical image registration, including modalities such as CT, CMR, and PET (Klein* et al., 2010). We chose Elastix for its flexibility in handling various registration tasks and the possibility of choosing different frameworks easily by modifying the configuration parameters file.

In our experiments, we employed a multi-resolution strategy to improve the registration process's accuracy and robustness. We utilized the MI metric. For the transformation model, we focused on rigid transformations, ensuring that the registrations accounted for rotations and translations without altering the shape of the structures.

The optimization was performed using the Adaptive Stochastic Gradient Descent (ASGD) optimizer, which provides efficient convergence with a maximum of 2000 iterations per resolution level. Additionally, we experimented with various parameter settings from the Elastix Model Zoo, (Klein* et al., 2010) to fine-tune the registration process for our specific use case. We tested different configurations by alternating which image (CTA or CMR) was designated as the fixed or moving image, to evaluate the impact on registration performance.

As a second experiment, we also attempted to register the segmentation masks directly, treating them as binary images. In this case, we use MSE as the similarity metric, calculated as:

$$\text{MSE} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \left[ I(i, j) - K(i, j) \right]^2 \qquad (1)$$

where: $I$ is the original image and $K$ is the reconstructed image.

### *3.4. Generation*

Diffusion Models usually consist of two stages:

1. A forward stage to gradually adding noise
2. A reverse stage to denoise and recover an original sample step-by-step

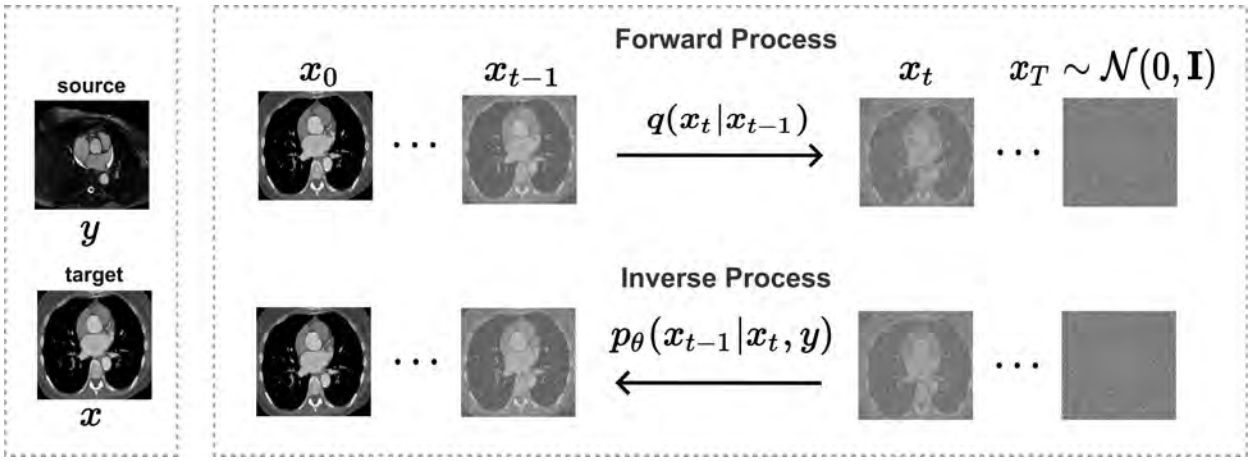



Figure 6: The forward diffusion process $q$ (left to right) gradually adds Gaussian noise to the target image. The reverse inference process $p$ (right to left) iteratively denoises the target image conditioned on a source image $y$. Source image $x$ is the CMR image

Representative frameworks include denoising diffusion probabilistic models (DDPM) (Ho et al., 2020), noise-conditioned score networks (NCSN) (Yang et al., 2021), and stochastic differential equations (Song and Ermon, 2019).

### *3.4.1. Denoising Diffusion Probabilistic Models*

DDPM has its diffusion stage including multiple small steps. In each step, a data sample is slightly corrupted by Gaussian noise. In DDPM the forward process is defined as a Markovian process. Gaussian noise is added in successive steps to obtain a set of noisy samples. In its reverse stage, DDPM performs a denoising task to recover an original image. where each step is also a Gaussian distribution. Then a neural network is trained to approximate each reverse diffusion step and estimate the mean and the covariance.

In this model, $x_0$ represents an original image and $q(x_0)$ denotes the original distribution of $x_0$, where $x_0 \sim q(x_0)$. A sequence of gradually corrupted images $x_1, x_2, \ldots, x_T$ are obtained according to the following Markovian process:

The forward process gradually adds Gaussian noise to an image over multiple steps:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t;\ \sqrt{1-\beta_t} \cdot x_{t-1}, \beta_t \cdot I) \tag{2}$$

$$q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1}), \tag{3}$$

where $T$ is the number of diffusion steps. The variances $\beta_t$ are selected such that the chain converges to a normal Gaussian distribution at step $T$, $q(x_T) \simeq \mathcal{N}(x_T; 0, \mathbf{I})$.

To avoid calculating all the intermediate steps, a closed-form expression is:

$$q(x_t \mid x_0) = \mathcal{N}(x_t;\ \sqrt{\bar{\alpha}_t} \cdot x_0, (1-\bar{\alpha}_t) \cdot \mathbf{I}). \tag{4}$$

Where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$

The reverse process aims to recover the original image by denoising the noisy images step-by-step. We can train a neural network $p_\theta(x_{t-1}|x_t)$ that receives as input the noisy image $x_t$ and the embedding at time step $t$, and learns to predict the mean $\mu_\theta(x_t, t)$ and the covariance $\Sigma_\theta(x_t, t)$:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{5}$$

The reverse step is conditioned on $x_0$ and $x_t$:

$$q(x_{t-1} \mid x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}(x_t, x_0), \tilde{\beta}_t \cdot \mathbf{I}), \tag{6}$$

To simplify the objective function, the mean and variances in are reformulated as:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\bar{\alpha}_t}} \left( x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right), \tag{7}$$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \beta_t. \tag{8}$$

The model is trained to minimize the following objective:

$$\mathcal{L}_t^{\text{simple}} = \mathbb{E}_{t, x_0, \epsilon_t} \left[ \left\| \epsilon_t - \epsilon_\theta \left( \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t} \epsilon_t, t \right) \right\|^2 \right] \tag{9}$$

### *3.5. Stochastic Differential Equations (SDE)*

The stochastic differential equation (SDE) approach also gradually transforms the original data distribution into a Gaussian distribution in the forward stage. Unlike the SDE method handles a continuous process.

The reverse process of this diffusion can be modeled with a reverse-time SDE which requires the score function of the density at each time step. Therefore, employs a neural network to estimate the score functions, and generates samples from $p(x0)$ by employing numerical

SDE solvers. The forward SDE that describes the process of transforming data into a Gaussian distribution:

$$\partial x = f(x,t)\cdot\partial t + \sigma(t)\cdot\partial\omega, \tag{10}$$

where $\omega_t$ is Gaussian noise, $f$ is a function of $x$ and $t$ that computes the drift coefficient, and $\sigma$ is a time-dependent function that computes the diffusion coefficient. The reverse SDE is used to recover the original data distribution:

$$\partial x = \Big[f(x,t) + \sigma(t)^2\cdot\nabla_x \log p_t(x)\Big]\cdot\partial t + \sigma(t)\cdot\partial\hat{\omega}, \tag{11}$$

where $\hat{\omega}$ represents the Brownian motion when the time is reversed, from $T$ to 0. We can train the neural network $s_\theta(x,t)\approx\nabla_x \log p_t(x)$ by optimizing:

$$\mathcal{L}^*_{dsm} = \mathbb{E}_t\Big[\lambda(t)\mathbb{E}_{p(x_0)}\mathbb{E}_{p_t(x_t|x_0)}\|s_\theta(x_t,t) - \nabla_x \log p_t(x_t|x_0)\|_2^2\Big], \tag{12}$$

### 3.5.1. Conditional DDPM

Saharia et al. (2021) proposed the conditional DDPM. Given the co-registered CTA and CMR pairs $(x^i, y^i)_{i=1}^K$, where $K$ is the number of image pairs, our objective function of eq. 9 is as follows:

$$\mathcal{L}_t^{\text{simple}} = \mathbb{E}_{t\sim[1,T],x_0,\epsilon_t}\left[\left\|\epsilon_t - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0^i + \sqrt{1-\bar{\alpha}_t}\epsilon_t, y, t\right)\right\|^2\right]. \tag{13}$$

The sampling process is a reverse Markovian process starting from a Gaussian noise $x_T\sim\mathcal{N}(0,I)$, in this case the reverse process is modified as:

$$q(x_{t-1}|x_t,x_0,y) = \mathcal{N}(x_{t-1};\tilde{\mu}(x_t,x_0,y),\tilde{\beta}_t\cdot I), \tag{14}$$

where

$$\tilde{\mu}_\theta(x_t,y,t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t,y,t)\right). \tag{15}$$

As illustrated in Figure 6, the CMR image is our condition $y$, and $x_0$ is our CTA image. UNet, (Ronneberger et al., 2015), was adopted for the reverse diffusion process to denoise.

### 3.5.2. Conditional SDE

The reverse-time SDE should be solved under the guidance of a condition of interest, in our case the CMR image as the condition. To supervise the forward and backward diffusion processes. Eq. 10 can be changed to:

$$dx = \sigma^t\cdot d\omega, \tag{16}$$

$$p_{0t}(x(t)|x(0)) = \mathcal{N}(x(t);x(0),\sigma(t)\cdot I), \tag{17}$$

where $t\sim\mathcal{U}([0,T])$. As $p_{0t}(x(t)|x(0))$ is a Gaussian perturbation kernel, the gradient of the perturbation kernel is

$$\nabla_{x(t)}p_{0t}(x(t)|x(0)) = -\frac{x(t)-x(0)}{\sigma(t)}$$

The objective function becomes

$$\mathcal{L}^*_{\text{dsm}} = \mathbb{E}_t\left[\lambda(t)\mathbb{E}_{x(0)}\mathbb{E}_{x(t)|x(0)}\left\|s_\theta(x(t),y,t) + \frac{x(t)-x(0)}{\sigma(t)}\right\|_2^2\right]. \tag{18}$$

The reverse-time SDE can be expressed as

$$dx = -\sigma^{2t}s_\theta(x(t),y,t)dt + \sigma^t d\bar{\omega}. \tag{19}$$

To sample from the time-dependent score-based model $s_\theta(x(t),y,t)$, we first draw a sample $x_T$ from the prior distribution $p_T\sim\mathcal{N}(x(0),\sigma(T)\cdot\mathbf{I})$, and then solve the reverse-time SDE numerically.

Three sampling techniques were used, Euler-Maruyama (EM), Prediction-Corrector (PC), and probability flow ordinary differential equation (ODE).

*Euler-Maruyama method*

To solve the reverse-time SDE, a simple discretization strategy is adopted, replacing $dt$ with a small increment $\Delta t$ and $d\bar{\omega}$ with a Gaussian noise $z\sim\mathcal{N}(0,\Delta t\cdot\mathbf{I})$. Then, we have

$$x_{t-\Delta t} = x_t + \sigma^{2t}s_\theta(x(t),y,t)\Delta t + \sigma^t\sqrt{\Delta t}z_t, \tag{20}$$

where $z_t\sim\mathcal{N}(0,\mathbf{I})$.

*Prediction-Correction method*

The PC sampling alternates between prediction and correction steps. The predictor can be any numerical solver for the reverse-time SDE with a fixed discretization strategy, such as the EM method. The corrector can be any score-based Markov Chain Monte Carlo method, such as Langevin dynamics. To implement it, it is necessary to calculate a Langevin step size $\gamma$:

$$\gamma = \left(\frac{r\|z\|_2}{\|s_\theta(x_i,y,\sigma_i)\|_2}\right)^2, \tag{21}$$

where $r$ is a signal-to-noise ratio, and $z\sim\mathcal{N}(0,\mathbf{I})$. Once the Langevin step size $\gamma$ is determined, we can sample according to Langevin dynamics.

*Probability flow ODE method (ODE)*

For any SDE in the form of eq. 10, there exists an associated ODE

$$dx = \left[f(x,t) - \frac{1}{2}g(t)^2\nabla_x \log p_t(x)\right]dt, \tag{22}$$

which has the same marginal probability density $p_t(x)$ trajectory as that of the SDE. Sampling by solving the reverse-time SDE is equivalent to solving eq. 22 in the reverse time direction. ODE sampling process starts from obtaining $x_T$ from $p_T$. Then, we integrate ODE in the reverse time direction and finally get a sample from $p_0$. In this case, the ODE equation is written as follows:

$$dx = -\frac{1}{2}\sigma^{2t}s_\theta(x(t),y,t)dt. \tag{23}$$

### 3.6. Dataset Configurations for Enhanced Training

In this study we explore different data modifications to reduce the data variability and focus the training in the anatomical differences between modalities. These modifications are outlined below:

- *Cropping*: We cropped all images to a fixed size of 256x256 pixels. Focusing on the aortic valve.
- *Image Intensities Inverted*: It helps the training by aligning the intensity distributions of the two modalities, as MRI and CT typically have opposite intensity representations. This process reduces the modality gap, making it easier for the model to learn the accurate mapping between
- *Removing Slices*: post-registration CMR and CTA datasets often contain slices with little to no useful information, especially in the last ones. We removed slices when more than 60% of the image was black.
- *Hounsfield Units Small Range*: For CTA images, Hounsfield Units (HU) are used to represent the density of tissues. We can emphasize specific tissue types or structures of interest by limiting the range of Hounsfield Units. This is important in the used model as it learns from the voxel intensities.

### 3.7. Training

In the training phase, we utilized the U-Net model proposed by Saharia et al. (2021). The parameters to configure included: batch size of 4, employing the Adam optimizer and learning rate of 1e-4. The image input size was standardized to 512x512 pixels, and the number of diffusion steps to 1000. For conditioned score-matching diffusion models: batch size of 4, employing the Adam optimizer, learning rate of 1e-4. Additionally, both diffusion and sampling steps were set to 1000. We utilized ODE, Euler, and PC samplers for generating samples from each trained model. To monitor performance during training, four sample images were generated every 50 epochs.

#### 3.7.1. Repeatability Assessment

We repeated sample generation five times to ensure result robustness and assess consistency. Following this, we calculated the mean values and standard deviation of pixel data, resulting in two new volumes for each model.

### 3.8. Quantitative Assessment

Generative models are expected to have two main characteristics: generation diversity and fidelity to the original dataset. There exist metrics to quantitatively assess these characteristics. Two commonly used metrics for this purpose are the Structural Similarity Index Metric (SSIM) and Peak Signal-to-Noise Ratio (PSNR).

#### 3.8.1. Structural Similarity Index Metric

SSIM is a perceptual metric that quantifies image quality degradation caused by processing such as data compression or transmission losses. It measures the similarity between two images. The SSIM value ranges from -1 to 1, where 1 indicates perfect structural similarity.It is calculated by Eq. 24.

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{24}$$

where: $\mu_x$ and $\mu_y$ are the average values of images $x$ and $y$. $\sigma_x^2$ and $\sigma_y^2$ are the variances of $x$ and $y$. $\sigma_{xy}$ is the covariance of $x$ and $y$. $C_1$ and $C_2$ are constants to stabilize the division with weak denominator.

If a pair of synthetic images are sampled, a low SSIM value would mean that the compared images are not structurally similar and, therefore, implies diversity.

#### 3.8.2. Peak Signal-to-Noise Ratio (PSNR)

Fidelity to the original dataset can be assessed using the Peak Signal-to-Noise Ratio (PSNR). PSNR measures the quality of reconstruction of lossy compression codecs. The signal in this case is the original data, and the noise is the error introduced by compression. The PSNR value is expressed in decibels (dB).

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}_I^2}{\mathrm{MSE}}\right) \tag{25}$$

where:

- $\mathrm{MAX}_I$ is the maximum possible pixel value of the image.
- MSE is the mean squared error between the original and the reconstructed image.

A high PSNR value indicates that the reconstructed image is high quality and closely resembles the original image.

## 4. Results

### 4.1. Aorta Segmentation

In Tables 1 and 2 the quantitative segmentation results obtained on the aorta CTA and CMR validation set are presented. The 3D-UNet model achieved the highest DSC and NSD values, indicating superior segmentation accuracy and surface agreement compared to 2D-UNet. In the second Table, TotalSeg model achieved the highest DSC and AUC values, while the 3D-UNet model showed the best performance in terms of NSD, indicating a closer surface match.

To have a qualitative result, Figure 7 shows an example of the CTA aorta segmentation result using TotalSegmentator.

Table 1: Dice Score (DSC), Normalized Surface Distance (NSD), and Area Under the Curve (AUC) for CMR-based segmentation.

| | **DSC** | **NSD** | **AUC** |
|---|---|---|---|
| 3D-UNet | 0.987±0.005 | 0.999±0.001 | 0.992±0.004 |
| 2D-UNet | 0.976±0.009 | 0.994±0.003 | 0.985±0.010 |

Table 2: Dice Score (DSC), Normalized Surface Distance (NSD), and Area Under the Curve (AUC) for CTA-based segmentation.

| | **DSC** | **NSD** | **AUC** |
|---|---|---|---|
| TotalSeg | 0.980±0.005 | 0.988±0.001 | 0.982±0.005 |
| 3D-UNet | 0.971±0.002 | 0.996±0.002 | 0.984±0.003 |
| 2D-UNet | 0.970±0.003 | 0.996±0.002 | 0.984±0.002 |

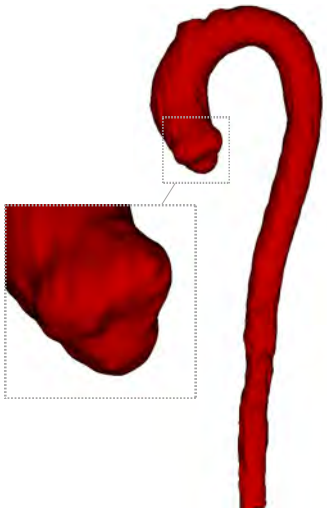


Figure 7: Example CTA Aorta Segmentation

Additionally, for qualitative results of the CMR,, Figure 8 shows an example of the CMR aorta segmentation. Figure 8a presents the frontal view of the segmented aorta, while 8b provides an inferior view of the aortic valve. However, it is important to note that while the aorta itself is well-segmented, the three cusps of the aortic valve are not precisely defined.

(a) (b)

Figure 8: Example CMR Aorta Segmentation

### *4.2. Registration*

#### *4.2.1. Deep Learning based*

The quantitative results of the registration in the validation set for the different loss functions experimented are presented in Table 3 and Table 4. The first Table presents the results using MI as the similarity loss, and Table 4 shows the results when using MIND as the similarity loss. It is shown that both metrics provide good performance in the registration with the TransMorph model, with Dice Scores values above 0.95 in almost all the cases. In the case of VoxelMorph the performance was very low.

Table 3: Registration Results using MI loss

| Model | MI loss | DSC loss | DSC |
|---|---|---|---|
| TransMorph | 0.5 | 0.5 | 0.9524 |
| | 1 | 0 | 0.5836 |
| | 1 | 0.5 | 0.9506 |
| | 1 | 1 | 0.9557 |
| VoxelMorph | 1 | 1 | 0.5965 |

Table 4: Registration Results using MIND

| Model | MIND loss | DSC loss | DSC |
|---|---|---|---|
| TransMorph | 0.5 | 0.5 | 0.9491 |
| | 1 | 0 | 0.5554 |
| | 1 | 0.2 | 0.9374 |
| | 1 | 1 | 0.9537 |
| VoxelMorph | 1 | 1 | 0.5820 |

A qualitative example of the deep-learning results is shown in Figure 9. Pre and post-registration, respectively, where the orange arrows show the aortic valve. In this registration approach, dice scores were very high; unfortunately, during a quick qualitative assessment, we saw that the deformable registration results were not useful for our purpose because the aortic valve needed to maintain its shape, and in some cases, it disappeared.

#### *4.2.2. Traditional Registration Methods*

The registration quality was evaluated using Dice Score, comparing the alignment between the segmentation masks. We compared results by switching the fixed and moving images: CTA as fixed with CMR moving, and vice versa. Figure 10 illustrates the Dice Score results for the registration using the MI similarity metric. Three different settings were evaluated: CMR Fixed (No Mask), CMR as Fixed, and CTA as Fixed image. Figure 11 presents the Dice Score results for registration using the MSE similarity metric. Two different settings were evaluated: CTA Fixed and CTA Fixed >87.This last, represents the registered cases with a Dice score

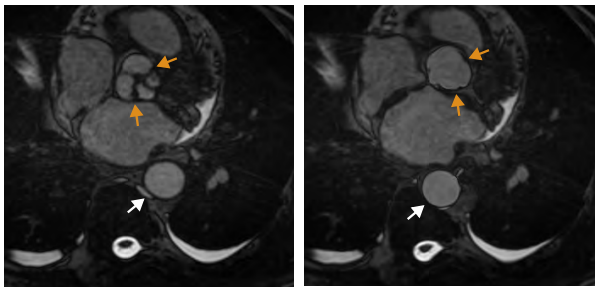


Figure 9: Example of the deep learning-based registration using the segmentation masks ans TransMorph model. Image before (left) and after (right) registration, were yellow arrows are pointing at the aortic valve.

value above 0.87. Figures 12a and 12b show examples of CMR aorta registration using MI with CTA fixed and CMR fixed settings, respectively. Figure 13 shows an example of CMR aorta registration using MSE as the similarity metric. Additionally, Figure 14 demonstrates the results of multimodal registration using traditional methods. The Figure shows pairs of CMR (top row) and CTA (bottom row) images, in this case when using the MSE as the similarity metric and CTA as the fixed image.

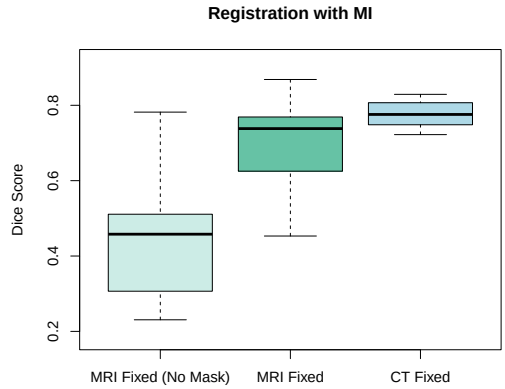



Figure 10: Dice Score results for registration using MI

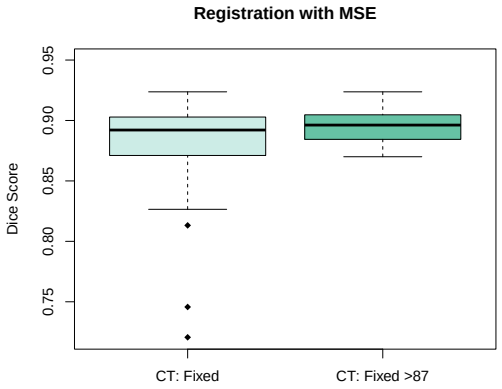



Figure 11: Dice Score results for registration using MSE

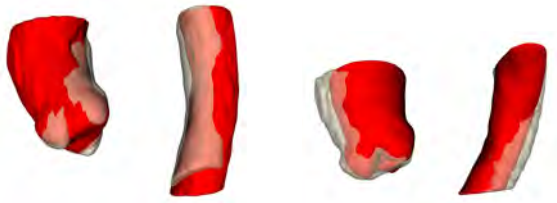



Figure 12: Example of Aorta Registration using MI where: a) CTA is used as moving and CMR as fixed, b) CTA is used as fixed and CMR as moving. Volume in color red represents the fixed volume and white the moving volume.

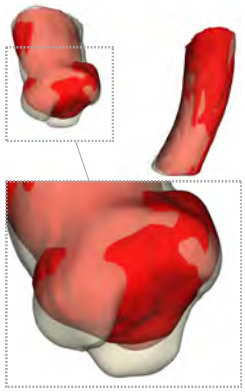


Figure 13: Example CMR Aorta Registration using MSE. The beige color represents the CMR segmentation volume, while the red color represents the CTA segmentation volume.

### *4.3. Diffusion Models*

#### *4.3.1. Dataset Configurations*

The final variations in the dataset included: standardizing the Hounsfield Units in the CTA, and removing black slices from the volumes. Although cropping and inverting image intensities were experimented with, they were not included in the final configuration. Examples of these results are shown in the Appendix A.

#### *4.3.2. Qualitative results*

Figure 15 and Figure 16 show examples of the image conversion using DDPM and score-matching models. Figure 15 shows the comparison between the different models' sampling generation, it shows the source CMR image, and the target CTA image, all the methods when sampling one time. Figure 16 represents the image generation across the four methods when sampling five times and obtaining the mean of the results. Here, it is also shown the source image: CMR, and the target: CTA.

Additionally, for having a global representation of the robustness performance, Figure 17 demonstrates the repeatability of an image sample by generating five samples using the same target CMR image. Here, it is also shown the standard deviation across all the samples,

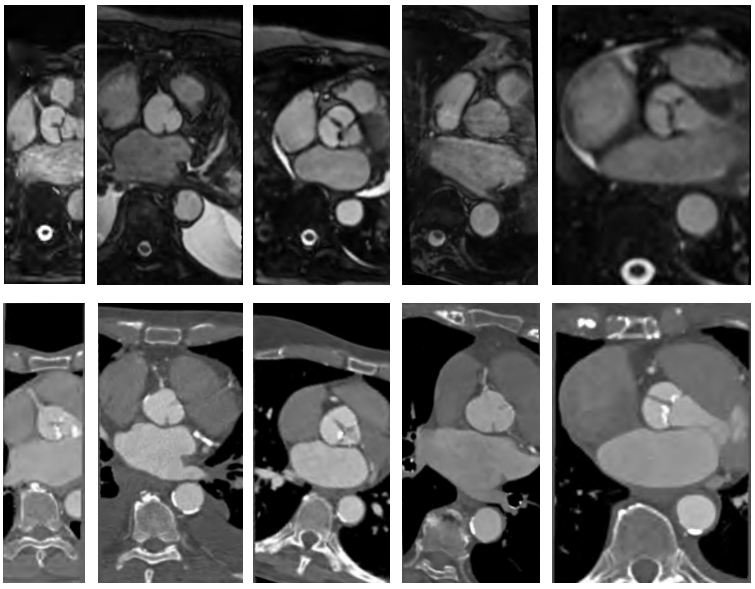


Figure 14: Example of multimodal registration. The figure displays pairs of CMR (top row) and CTA (bottom row) images. Each pair represents the alignment of anatomical structures between the two modalities.

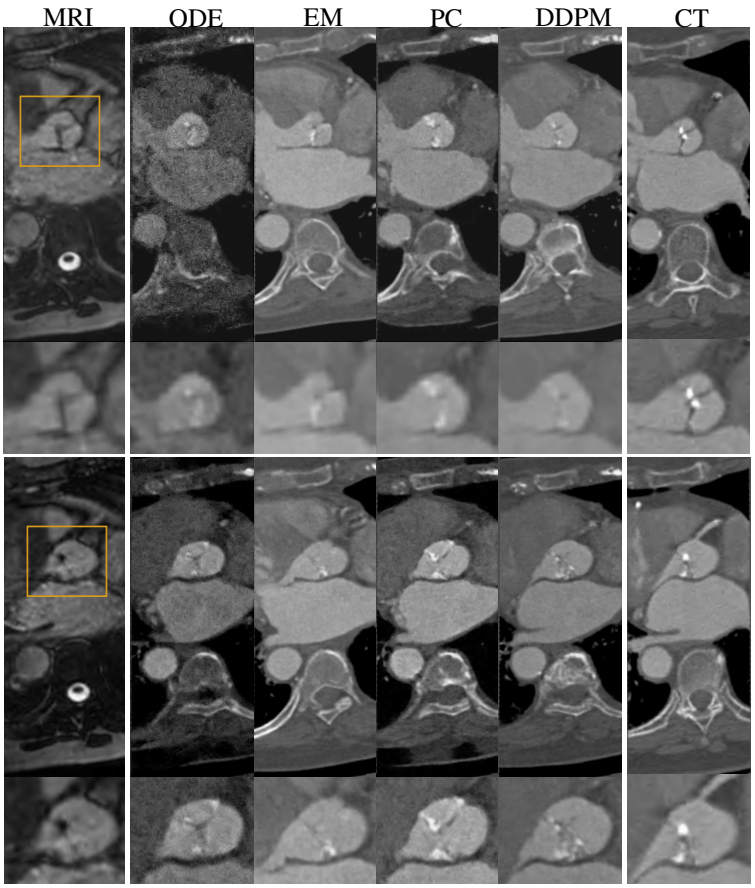



Figure 15: Comparison of different diffusion models' outputs. The Figure showcases the generated samples from each model, comparing the output with the original CMR (source) and CTA (target) images. Two example images are provided to illustrate the performance of the models.

to have a better understanding of where the model is adding more changes when generating. In this example, it was using the Score-matching model and PC as the sampling method.

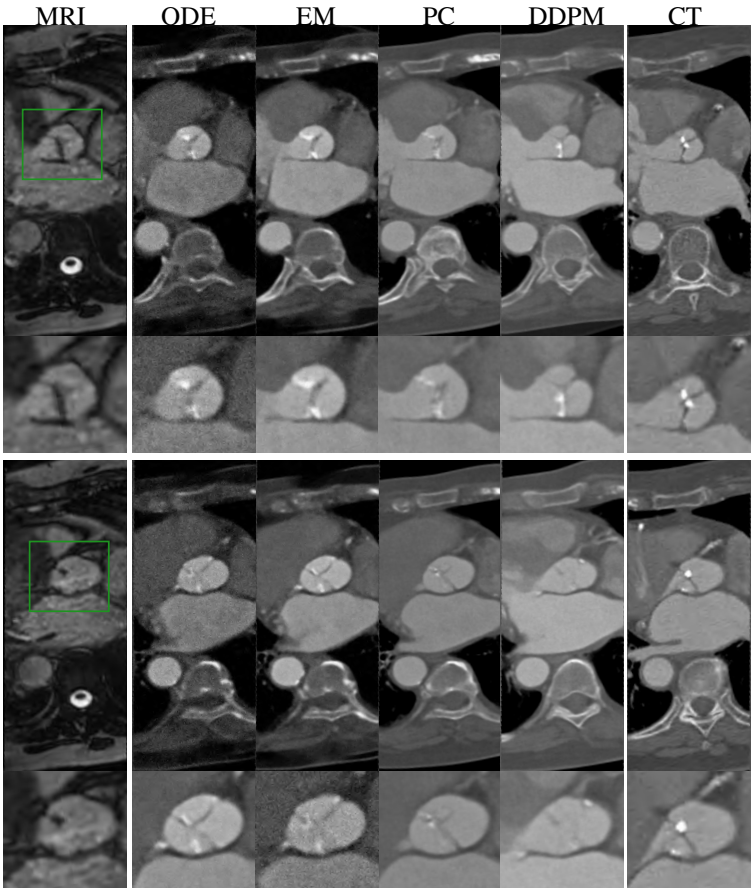



Figure 16: Comparison of different diffusion models' outputs in two example samples. The Figure displays the mean average of 5 generated samples for each model, comparing the output with the original CMR (source) and CTA (target) images.

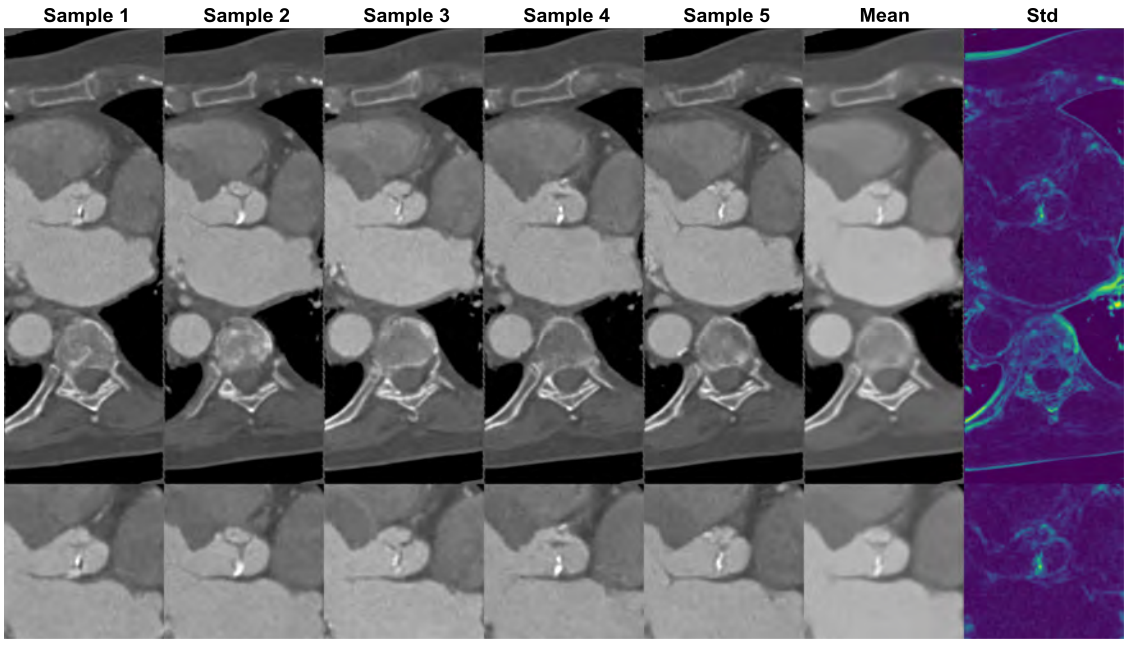



Figure 17: Example of repeatability. This Figure showcases five generated samples of a CTA scan using the Conditioned Score-Matching diffusion model and the PC sampling strategy. The Mean column displays the average of the five generated samples. The last column shows the standard deviation.

#### *4.3.3. Quantitatively assessment*

In this study, we utilized the SSIM and PSNR metrics to evaluate image quality. The mean and standard deviation of the SSIM and PSNR values among the validation images are shown in Figure 18a and 18b

Figure 18a shows the SSIM and PSNR values for the validation images. The results indicate that DDPM and PC methods consistently achieve higher SSIM and PSNR values. EM and ODE methods, demonstrate lower performance metrics in comparison.

To have a better idea of the performance of the generation we also compared the SSIM and PSNR values, when sampling the images five times. Figure 18b displays the average SSIM and PSNR results across five samples, providing a more comprehensive comparison of the methods' performance.

#### *4.3.4. Computational Time Analysis*

The training process for both models lasted 72 hours. An estimation of the average inference time for each model on a set of one axial slice shows that the ODE model was the fastest, taking 50 seconds. The DDPM-conditioned method required 120 seconds, while the EM method took 180 seconds. The score-matching model with the Predictor-Corrector (PC) sampler was the slowest, consuming 420 seconds due to the algorithm's prediction and correction steps.

## 5. Discussion

In this study, we investigated the application of diffusion model techniques to convert CMR to equivalent CTA images. Our pipeline involves data construction, segmentation, registration, and generation.

### *5.1. Segmentation*

The masks were generated using well-known approaches, TotalSegmentator and nnUnet. Both models showed good performance with the DICE score, all values above 0.97 compared with our ground truth, as shown in Tables 1 and 2. When comparing the visualization results in each modality, we observed that for CMR segmentation, the details in the segmentation are of high quality and the smoothness was well done. This observation can be attributed to the nnUnet architecture which applies preprocessing and postprocessing to the masks. On the other hand, CTA segmentation also yielded good results. For CTA the difference of using a pre-trained or a self-trained model was not statistically significatant. However, it is important to emphasize that there are still some irregularities between pixels; furthermore, unlike CMR images, the aortic valve is not precisely defined. We attribute this phenomenon to both interpolations by Total-Segmentator as well as the voxel space in CTA images but it can also be due to the difference in modality information..

### *5.2. Registration*

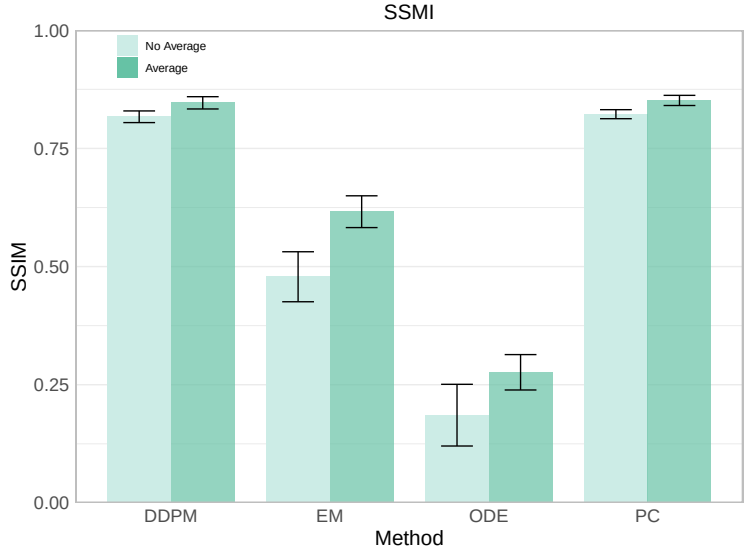
We proposed exploring deep-learning and non-deformable methods for multimodal registration between CTA and MRI. Two deep learning approaches were explored: VoxelMorph and TransMorph, which are state-of-the-art for registration. TransMorph demonstrates superior performance in DICE scores, while VoxelMorph produces low results. Although TransMorph tends to produce a high score, when visualizing the registration we concluded that it did not preserve the organ's anatomy, as observed in Figure 9, where the aortic valve changed its anatomy after registration. This phenomenon is attributed to the deformable regularizer parameter requiring further hyperparameter exploration among others. In the case of VoxelMorph, we can also attribute the low DICE to the absence of attention blocks present in Transmorph models adding more attention to segmentation.

To address this issue and explore alternatives for preserving anatomical features, traditional methods were considered. We experimented with fixed and moving images alongside similarity losses resulting in good registration results as shown in Figure 7. Notably using CTA as a fixed configuration yielded better results indicating its well-defined structures made it easier as a base image for registering.

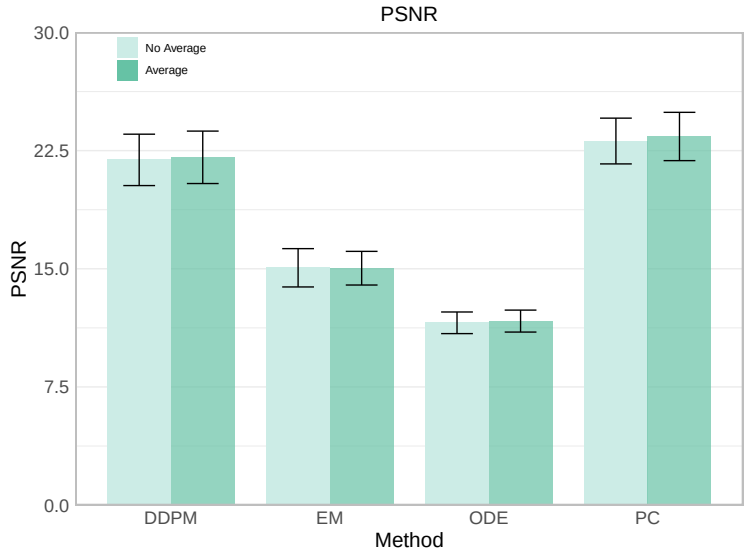
### *5.3. Generation*

We experimented with various training data set variations to compare the impact of different modifications. Adjusting the window in the Hounsfield field units and removing slices, refocused the model on structures with contrast agent, while also eliminating unimportant areas such as lungs for our application. These variations revealed that some improvements were made to the results, while others did not. This reflects how increasing variability in the training set influences what the model learns and directly impacts the final CTA image.

We explored two different diffusion models: DDPM and score matching. In the case of score matching, three different samplers were used. Comparing the two methods, we observed that DDPM achieved good results for SSMI and PSNR. However, the score-matching methods yielded low results when using ODE and EM samplers. This suggests a need for further exploration in the diffusion and sampling steps because although the forward process was similar, the final sampling differed greatly compared to PC. It is important to note that ODE gave the worst results with very noisy images in SSMI and PSNR values. Among all methods tested, PC and conditioned DDPM showed the best performance overall, especially regarding anatomical structures for CMR-CTA. As we observed the results were different between the samplers, we explored the repeatability between all the methods, in this case generating the same image five times, this experiment gave us some interesting findings, the smoothness of the image when there




(a)




(b)

Figure 18: Comparison of a) Structural Similarity Index Measure (SSIM) and b) Peak Signal-to-Noise Ratio (PSNR) across different methods: Denoising Diffusion Probabilistic Models (DDPM), Euler-Maruyama (EM), Ordinary Differential Equation (ODE), and Predictor-Corrector (PC) and the corresponding average method

is texture. Additionally, there was slight improvement in noise reduction. Furthermore, while usually having more variation in image details, it was particularly interesting to find this specially in the calcifications. These finding suggests that we need to enforce our model to learn details in the image, capturing the small characteristics, including global and local information when converting the image. Overall image performance is satisfactory but could be improve via hyper-parameter tuning. The selection of training data plays a crucial role along with registration since method learning can be influenced by even minor differences among training images resulting from exploring diverse data variables

### *5.4. Limitations*

Despite demonstrating promising results, our pipeline has certain limitations and areas for improvement. Firstly, the model encounters difficulties in replicating small details in the CTA, including calcifications.

Secondly, the diffusion models applied are very sensitive to image training; the data needs to be very similar in the training step, leading to the generation of images that are only similar to the CMR source. In some cases, the models produce inaccurate shapes, especially when there is motion, multiple organs, or many details in the image. This could also be related to the different images

used in the training that can vary significantly for each patient. Moreover, the registration and segmentation accuracy directly impact the image generation, as errors in both previous steps propagate through the pipeline. In addition, the variability in both modalities was a limiting factor, as some did not have consistent triggering and others exhibited a significant amount of cardiac motion. Furthermore, the year of acquisition significantly influenced the image quality.

## 6. Future work

To enhance the performance of the pipeline, future work could include:

- Incorporating local attention mechanisms to focus on calcifications.
- Optimizing hyper-parameters to improve the robustness of the models to handle variations in patient data more effectively.
- Extending the pipeline to generate the contrast agent using the CT as the source image, providing a valuable tool for cases where contrast agents are contraindicated.
- More segmentation and registration techniques could further improve the accuracy of image generation; for example Atlas-based registration.
- Cropping the aorta volume and train the different models. Removing other anatomical structures.
- Investigating the integration of additional clinical data could improve the model and extend its applicability across different patient demographics.

## 7. Conclusions

In this study, we have presented an end-to-end pipeline for generating CTA images using CMR as the source. The method is based on denoising diffusion probabilistic models (DDPMs) and score-matching models. We used CMR images as the condition for training and CTA images as the target.

Through a comprehensive evaluation, the results have shown promising outcomes both quantitatively and qualitatively for the CMR-to-CTA conversion. While the method sometimes overestimated calcifications in the CTA images, the overall structural similarity demonstrated promising results using the PC sampler and DDPM. We achieved promising results for PSNR (23.4) and SSMI (0.84). This indicates that the pipeline is capable of maintaining major anatomical structures accurately.

However, there are certain limitations and areas for further improvement. The model encountered difficulties in replicating small details in the CTA, such as calcifications. Additionally, the diffusion models are highly sensitive to the training data, requiring very similar images for effective training. In some cases, the models produced inaccurate shapes, especially when there was motion, multiple organs, or many details in the image. These issues could be related to the variability in the training images, which can differ significantly among patients.

## Acknowledgments

I would like to extend my deepest gratitude to my supervisors, for the guidance and feedback they gave me throughout this project. I am immensely grateful to Tyrol Klinik and the Medical University of Innsbruck for hosting this project and providing the necessary computational resources for its execution. I extend my sincere appreciation to the MAIA master consortium and the European Commission for granting me this opportunity and funding my education. Lastly, I am also grateful to my family and friends who, despite the physical distances, have always been my emotional support.

## References

Al-Najafi, S., Sanchez, F., Lerakis, S., 2016. The crucial role of cardiac imaging in transcatheter aortic valve replacement (TAVR): Pre- and post-procedural assessment. Curr. Treat. Options Cardiovasc. Med. 18, 70.

Balakrishnan, G., Zhao, A., Sabuncu, M., Guttag, J., Dalca, A.V., 2019. Voxelmorph: A learning framework for deformable medical image registration. IEEE TMI: Transactions on Medical Imaging 38, 1788–1800.

Barbanti, M., Yang, T.H., Rodès Cabau, J., Tamburino, C., Wood, D.A., Jilaihawi, H., Blanke, P., Makkar, R.R., Latib, A., Colombo, A., Tarantini, G., Raju, R., Binder, R.K., Nguyen, G., Freeman, M., Ribeiro, H.B., Kapadia, S., Min, J., Feuchtner, G., Gurtvich, R., Alqoofi, F., Pelletier, M., Ussia, G.P., Napodano, M., de Brito, Jr, F.S., Kodali, S., Norgaard, B.L., Hansson, N.C., Pache, G., Canovas, S.J., Zhang, H., Leon, M.B., Webb, J.G., Leipsic, J., 2013. Anatomical and procedural features associated with aortic root rupture during balloon-expandable transcatheter aortic valve replacement. Circulation 128, 244–253.

Bradbury, R., Vallis, K.A., Papiez, B.W., 2024. Paired diffusion: Generation of related, synthetic pet-ct-segmentation scans using linked denoising diffusion probabilistic models. arXiv URL: https://arxiv.org/abs/2403.14066.

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. doi:https://doi.org/10.1016/j.media.2022.102615.

Croitoru, F.A., Hondru, V., Ionescu, R.T., Shah, M., 2023. Diffusion models in vision: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 45, 10850–10869.

Dorjsembe, Z., Odonchimed, S., Xiao, F., 2022. Three-dimensional medical image synthesis with denoising diffusion probabilistic models, in: Medical Imaging with Deep Learning (MIDL) 2022. URL: https://openreview.net/pdf?id=Oz7lKWVh45H.

Fanning, J.P., Platts, D.G., Walters, D.L., Fraser, J.F., 2013. Transcatheter aortic valve implantation (tavi): valve design and evolution. International journal of cardiology 168, 1822–1831.

Graf, R., Schmitt, J., Schlaeger, S., Möller, H.K., Sideri-Lampretsa, V., Sekuboyina, A., Krieg, S.M., Wiestler, B., Menze, B., Rueckert, D., Kirschke, J.S., 2023. Denoising diffusion-based MRI to CT image translation enables automated spinal segmentation. Eur. Radiol. Exp. 7, 70.

Hantao, Z., Jiancheng, Y., Shouhong, W., Pascal, F., 2024. Lefusion: Synthesizing myocardial pathology on cardiac mri via lesion-focus diffusion models. arXiv URL: https://arxiv.org/abs/2403.14066.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models, in: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 6840–6851.

Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods 18, 203–211.

Jhaveri, K.D., Saratzis, A.N., Wanchoo, R., Sarafidis, P.A., 2017. Endovascular aneurysm repair (EVAR)– and transcatheter aortic valve replacement (TAVR)–associated acute kidney injury. Kidney Int. 91, 1312–1323.

Kazerouni, A., Aghdam, E.K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., Merhof, D., 2023. Medical diffusion: Denoising diffusion probabilistic models for 3d medical image generation. ArXiv URL: https://arxiv.org/abs/2211.03364.

Klein*, S., Staring*, M., Murphy, K., Viergever, M.A., Pluim, J.P., 2010. elastix: a toolbox for intensity-based medical image registration. IEEE Transactions on Medical Imaging 29, 196 – 205.

Lindroos, M., Kupari, M., Heikkilä, J., Tilvis, R., 1993. Prevalence of aortic valve abnormalities in the elderly: an echocardiographic study of a random population sample. Journal of the American College of Cardiology 21, 1220–1225.

Lukas, S., Christoph, K., Julia, D., Simone, G., Axel, B., Guy, F., Bernhard, M., Gudrun, Maria, F., Agnes, M., Michael, C.G., Nikolaos, B., 2022. Minireview: Transaortic transcatheter aortic valve implantation: Is there still an indication? Frontiers in Cardiovascular Medicine 9.

Luo, Y., Hu, W., 2024. Measurement-guided diffusion models for high-quality medical image synthesis. ArXiv URL: https://arxiv.org/abs/2305.18453.

Lyu, Q., Wang, G., 2022. Conversion between ct and mri images using diffusion and score-matching models. arXiv preprint arXiv:2209.12104 .

Mahon, C., Mohiaddin, R., 2021. The emerging applications of cardiovascular magnetic resonance imaging in transcatheter aortic valve implantation. Clinical Radiology 76, 73.e21–73.e37. doi:https://doi.org/10.1016/j.crad.2019.11.011.

Mayr, A., Klug, G., Reinstadler, S.J., Feistritzer, H.J., Reindl, M., Kremser, C., Kranewitter, C., Bonaros, N., Friedrich, G., Feuchtner, G., Metzler, B., 2018. Is MRI equivalent to CT in the guidance of TAVR? a pilot study. Eur. Radiol. 28, 4625–4634.

Pamminger, M., Klug, G., Kranewitter, C., Reindl, M., Reinstadler, S.J., Henninger, B., Tiller, C., Holzknecht, M., Kremser, C., Bauer, A., Jaschke, W., Metzler, B., Mayr, A., 2020. Non-contrast MRI protocol for TAVI guidance: quiescent-interval single-shot angiography in comparison with contrast-enhanced CT. Eur. Radiol. 30, 4847–4856.

Pan, S., Abouei, E., Wynne, J., Chang, C.W., Wang, T., Qiu, R.L.J., Li, Y., Peng, J., Roper, J., Patel, P., Yu, D.S., Mao, H., Yang, X., 2023a. Synthetic CT generation from MRI using 3D transformer-based denoising diffusion model. Med. Phys. .

Pan, S., Wang, T., Qiu, R.L.J., Axente, M., Chang, C.W., Peng, J., Patel, A.B., Shelton, J., Patel, S., Roper, J., Yang, X., 2023b. 2d medical image synthesis using transformer-based denoising diffusion probabilistic model. Physics in Medicine and Biology 68. URL: https://api.semanticscholar.org/CorpusID:257954358.

Ram, P., Mezue, K., Pressman, G., Rangaswami, J., 2017. Acute kidney injury post–transcatheter aortic valve replacement. Clin. Cardiol. 40, 1357–1362.

Ricardo, M.d.A., Karla, S.M., Joan, C.V., Robert, M., 2023. Mam-e: Mammographic synthetic image generation with diffusion models. arXiv URL: https://arxiv.org/abs/2311.07945. 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.

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

Saharia, C., Chan, W., Ng, C., Norouzi, M., Fleet, D.J., Salakhutdinov, R., 2021. Image super-resolution via iterative refinement. arXiv preprint arXiv:2104.07636 .

Serhal, A., Koktzoglou, I., Aouad, P., Carr, J.C., Giri, S., Morcos, O., Edelman, R.R., 2018. Cardiovascular magnetic resonance imaging of aorto-iliac and ilio-femoral vascular calcifications using proton density-weighted in-phase stack of stars. J. Cardiovasc. Magn. Reson. 20, 51.

Song, Y., Ermon, S., 2019. Generative modeling by estimating gradients of the data distribution, in: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.

Stojanovski, D., Hermida, U., Lamata, P., Beqiri, A., Gomez, A., 2023. Echo from noise: Synthetic ultrasound image generation using diffusion models for real image segmentation, in: Kainz, B., Noble, A., Schnabel, J., Khanal, B., Müller, J.P., Day, T. (Eds.), Simplifying Medical Ultrasound, Springer Nature Switzerland, Cham. pp. 34–43.

Txurio, M.S., Román, K.L.L., Marcos-Carrión, A., Castellote-Huguet, P., Santabárbara-Gómez, J.M., Oliver, I.M., Ballester, M.A.G., 2023. Diffusion models for realistic ct image generation, in: Chen, Y.W., Tanaka, S., Howlett, R.J., Jain, L.C. (Eds.), Innovation in Medicine and Healthcare, Springer Nature Singapore, Singapore. pp. 335–344.

Wang, T., Lei, Y., Fu, Y., Wynne, J.F., Curran, W.J., Liu, T., Yang, X.R., 2021. A review on medical imaging synthesis using deep learning and its clinical applications. Journal of Applied Clinical Medical Physics 22, 11–36. doi:10.1002/acm2.13121.

Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A., 2022. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiol. Artif. Intell. 9, 1–9.

Yang, S., Jascha, S.D., Diederik, P.K., Abhishek, K., Stefano, E., Ben, P., 2021. Score-based generative modeling through stochastic differential equations arXiv:2011.13456.

## Appendix A. Additional Results

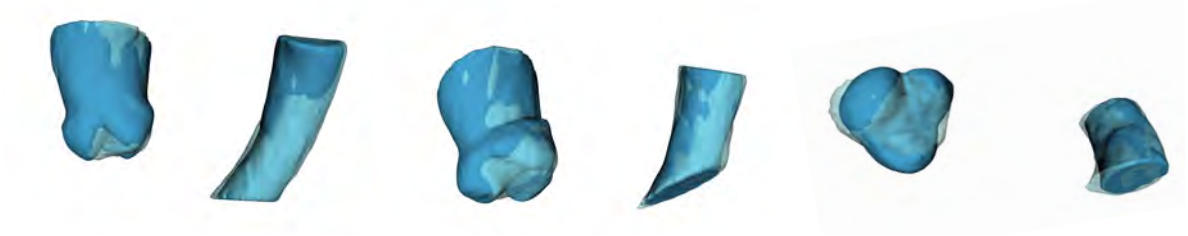


Figure A.19: Example of Multimodal Registration. The dark blue color represents the CMR segmentation volume, while the light blue represents the CTA segmentation volume.

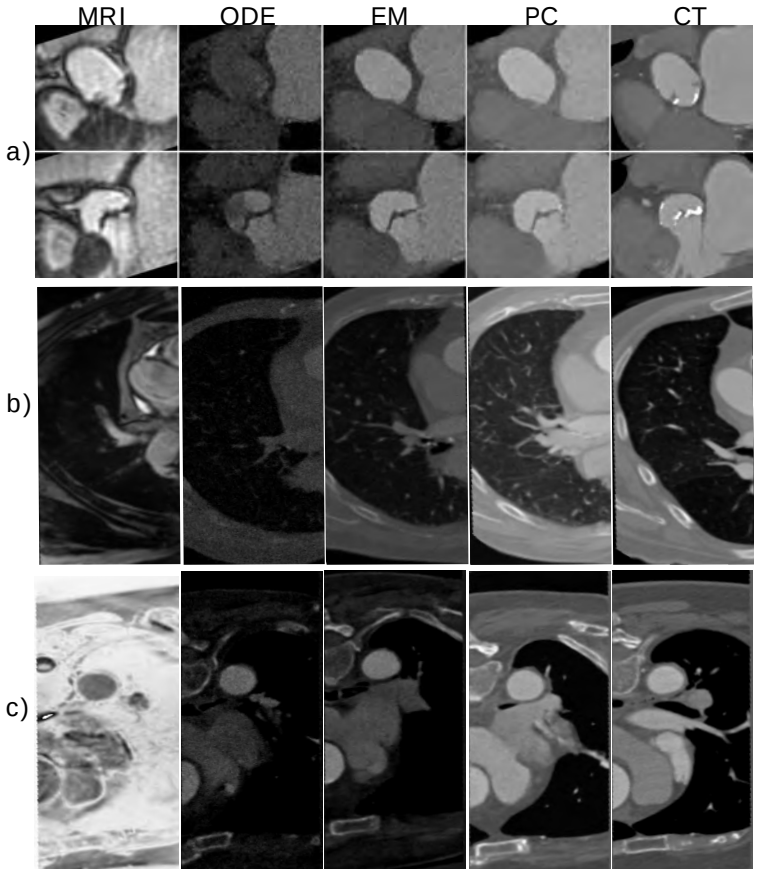
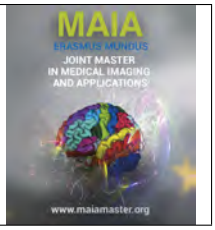



Figure A.20: Example of CMR-CTA conversion results when modifying the training dataset: a) by cropping the image, b) without altering the windowing of the HU, and c) by inverting the intensities of the MRI image.

Medical Imaging and Applications

Master Thesis, June 2024




# Adapting generalist vision language models for surgical phase recognition

Lisle Faray de Paiva, Kun Yuan, Vinkle Srivastav, Nicolas Padoy

*CAMMA lab / iCUBE - Universite de Strasbourg*

**Abstract**

This study presents a novel approach to enhancing surgical phase recognition by adapting generalist vision-language models, specifically the Surgical Vision Language Pre-training (SurgVLP) model. Our research leverages the SurgVLP model's contrastive pre-training on surgical lecture video-text pairs to integrate visual and textual data for improved phase recognition in laparoscopic cholecystectomy. We aim to merge visual and textual representations to boost model performance effectively by employing feature fusion techniques such as weighted sum and gated mechanisms. Our experiments reveal that the weighted sum method outperforms the gated mechanism, highlighting the effectiveness of directly combining image and text features. Additionally, we investigate the impact of text prompt generation methods, including hand-crafted prompts and those generated by large language models like GPT-4, on the accuracy and robustness of phase recognition. The results demonstrate that simple, well-constructed prompts can be as effective as those generated by advanced models.

*Keywords:* Vision-language models, Downstream tasks, Surgical phase recognition

## 1. Introduction

Cholecystectomy is the surgical removal of the gall bladder and the standard treatment of cholecystitis (Shabanzadeh et al., 2022), gallbladder inflammation. The procedure can be done with open surgery or laparoscopic techniques. However, laparoscopic cholecystectomy is the current standard technique as opposed to open surgery (Coccolini et al., 2015) due to being less invasive, causing less postoperative pain and a speedier recovery (Ziogas and Tsoulfas, 2017). During open cholecystectomy, the surgeon makes a 15 cm incision in the abdomen below the ribs on the right side to access the abdominal region (Coccolini et al., 2015). During laparoscopic cholecystectomy, instead of making an incision to access the cavity directly, a small 2 - 3 cm incision near the belly button is made for the placement of a laparoscope, and three additional small incisions in the upper right abdomen for insertion of the laparoscopic instruments exemplified in Figure 1, such as graspers for grasping the gallbladder, clip applier to control the cystic duct and artery, electrocautery for cutting and coagulating and a retrieval bag for the removal of the gallbladder from the cavity (Hassler et al., 2021).

The procedure is visualized through the video feed from the laparoscope, a small lighted camera. After the patient is prepped, the abdominal cavity is inflated using carbon dioxide, and trocars and tools are positioned as exemplified in Figure 1; the gallbladder is grasped and retracted over the liver to allow visualization of Calot's triangle (Olsen, 1991). Dissection is performed to achieve the critical view of safety, which requires three criteria to be met: (1) the hepatocystic triangle is clear of all fat and fibrous tissue; (2) the lower $\frac{1}{3}$ of the gallbladder is separated from the liver bed; and (3) only the cystic duct and cystic artery are connected to the gallbladder (Majumder et al., 2020). After this view is achieved, both structures are carefully clipped and transected, and then the gallbladder dissection off the liver bed is performed (Olsen, 1991). Once the dissection is complete, the area is checked for any bleeding that may have occurred, and the tissues and aspirate dry all fluids (Olsen, 1991). After aspirating all fluid and checking for hemostasis by slightly deflating the abdominal cavity, the gallbladder is placed inside a specimen pouch and removed from the abdomen. Once the gallbladder is removed, all trocars are removed, and port sites are closed (Hassler et al., 2021).

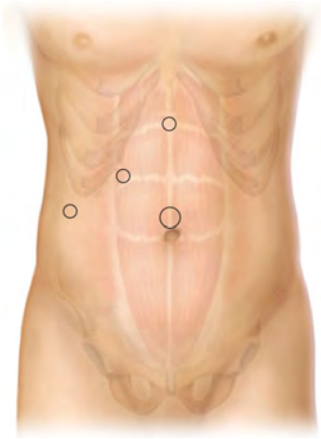


Figure 1: Port positions during laparoscopic cholecystectomy (Majumder et al., 2020).

Some downsides of laparoscopic surgery are the lack of hand tactile feedback, restricted view, and limited movement of laparoscopic instruments as well as a steeper learning curve and longer operative time (Buskens et al., 2014) (Han et al., 2015). As a way to improve surgical safety and efficacy during surgery, computer-aided surgery systems are developed. These procedures generate numerous surgical videos, creating new opportunities for applications in Computer Vision and Artificial Intelligence (AI). Analyzing these surgical videos can positively impact real-time performance (Hashimoto et al., 2018) with decision support tools that allow pre- and intra-operation information and insight into the surgical workflow (Kitaguchi et al., 2020). One such application is surgical phase recognition.

AI systems that can identify the surgical phases can benefit many tasks, such as education, evaluation of surgical performances, quality control, and analysis of complications (Golany et al., 2022). Standard practice involves building fully supervised deep-learning methods to recognize surgical phases, such as in Czempiel et al. (2020), which trains a Multi-stage Temporal Convolutional Neural Network for this specific task. These fully-supervised methods require extensive collaboration with clinical experts to annotate ground truth and validate it. They are typically trained on procedure-specific surgical video datasets from a single center, limiting their generalization and ability to represent the complexities of surgical workflows (Eisenmann et al., 2022). Additionally, multi-modal approaches integrate visual and textual data and can enhance phase recognition by leveraging complementary information from different data sources. These dependencies and limitations motivate us to utilize the emerging vision-language foundation models to address these issues.

Vision-language foundation models are massively scaled models trained on large amounts of data to be adapted to various downstream tasks, allowing them to develop unparalleled generalistic intelligence (Lam and Qiu, 2024). Examples of these models are the Contrastive Learning Image Pretraining (CLIP) model (Radford et al., 2021) developed by OpenAI; the CLIP model consists of a vision-language model pre-trained on 400 million image-text pairs obtained from the internet. Similarly, the Google ALIGN (Jia et al., 2021) model is trained on over one billion noisy image alt-text pairs without expensive filtering or post-processing steps. However, a domain gap exists between the natural image-text pairs used to train CLIP and ALIGN and those in the surgical domain. Thus, a vision-language foundation model pre-trained specifically on surgical data is necessary. The Surgvlp (Yuan et al., 2023) is a CLIP-like model, pre-trained on 1326 surgical video lectures sourced from WebSurg, EAES, and Youtube. These models have above-average zero-shot transfer capabilities; however, some downstream tasks, such as surgical phase recognition, require more adaptation to be functional in a clinical setting (Zhou et al., 2022b).

This research aims to enhance the safety, efficiency, and outcomes of surgical procedures by developing and adapting generalist vision-language models for surgical phase recognition. The ultimate goal is to create AI systems that are accurate, reliable, and capable of being easily adapted to various surgical tasks, thereby contributing to the broader field of computer-aided surgery and improving patient care. The primary objective of this thesis is to develop and adapt generalist vision-language models, specifically the SurgVLP model, for surgical phase recognition in laparoscopic cholecystectomy by partially fine-tuning the model while integrating visual and textual data through a feature fusion technique. The experiments demonstrated that the SurgVLP model, which incorporates both visual and textual information, consistently outperformed the ResNet50 vision-only model. In this work, we make the following contributions:

1. We explore and implement feature fusion techniques, such as weighted sum and gated mechanisms, to effectively combine visual and textual data for enhanced model performance;
2. We evaluate the impact of text prompt generation methods, including hand-crafted prompts and those generated by large language models, on the accuracy and robustness of phase recognition;
3. We assess the model's performance with different amounts of training data, thereby demonstrating its capability for few-shot learning and generalization;
4. We further investigate integrating additional textual information from medical literature to improve the model's contextual understanding and classification accuracy.

## 2. State of the art

This section reviews the literature on surgical computer vision, vision-language models, and prompt tun-

ing techniques.

### 2.1. Surgical Computer Vision

Surgical Computer Vision is a domain-specific subarea of Computer Vision that focuses on developing tools for analyzing surgical visual data. Surgical phase recognition is a domain-specific task that differs from general computer vision by requiring an understanding of the sequential flow and fine-grained details of surgical procedures to recognize surgical activities and objects accurately. Early work in this field concentrated on automatically recognizing surgical workflow through two primary tasks: phase recognition and tool presence detection. These tasks have been extensively studied using fully supervised models across various surgeries, including cataract, neurological, and laparoscopic procedures. Early approaches, such as in Padoy et al. (2012), relied on hand-crafted visual features and manually annotated tool usage signals. Recently, the rise of deep learning has introduced models that automatically learn features from surgical videos, enhancing accuracy and efficiency.

One notable development by Twinanda et al. (2016) is the EndoNet architecture, a fully supervised convolutional neural network (CNN) designed to perform phase recognition and tool presence detection simultaneously. EndoNet utilizes visual information exclusively, eliminating the need for additional equipment or manual annotations, and has achieved state-of-the-art results in these tasks.

Another significant advancement is the Multi-Stage Temporal Convolutional Network (MS-TCN) proposed by Czempiel et al. (2020), which performs hierarchical prediction refinement using causal, dilated convolutions. This fully supervised spatial-temporal model ensures smooth and accurate predictions during ambiguous transitions and has outperformed various Long Short-Term Memory (LSTM) based methods on laparoscopic cholecystectomy video datasets, both with and without additional tool information.

### 2.2. Vision-language Foundation models

The CLIP model proposed by Radford et al. (2021) uses natural language supervision to learn image representation. The model architecture, exemplified in Figure 2, explores two architectures: the ResNet50 and Vision Transformer (ViT) as the image backbone. They employ a Transformer with a base size of 63M parameters, 12 layers, and 512 wide with 8 attention heads for the text backbone. The model is trained from scratch with no pre-trained weights. The CLIP model requires enormous amounts of data to train on. However, manual annotation is cumbersome, expensive, and unpractical at this scale. To supply the required amount of data to train a foundation model, they sourced 400 million image-text pairs from the internet. Its pre-training

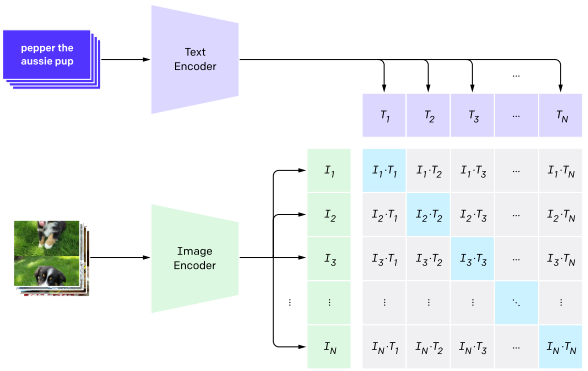



Figure 2: CLIP model architecture (Radford et al., 2021)

strategy focuses on predicting which text as a whole is paired with which image instead of trying to predict which exact words of the text accompany each image. During the pre-training, it maximizes the cosine similarity between the image and text embeddings of real pairs while minimizing the cosine similarity of incorrect pairings. Jointly training both backbones aligns the image and text representations to the multi-modal representation space. It uses a linear projection to map the embeddings from the encoder's representation space to the multi-modal representation space.

Shifting from images to video, Miech et al. (2019) explores using instructional videos and captions to pretrain a joint multi-modal embedding space. The dataset developed, Howto100M, comprises 136.6M video clips extracted from 1.22M videos sourced from YouTube with their corresponding captions automatically generated by the YouTube Automatic Speech Recognition(ASR) system. Their training strategy consists of extracting the video features at a frame-level and video level; they extract 2D frame-level features using the ImageNet pre-trained ResNet-152 at one frame per second and 3D video-level features using the Kinetics pretrained ResNeXt-101 16-frames model at 1.5 features per second. These features are aggregated through temporal max-pooling and concatenation, forming a 4096-dimensional vector for each video clip. For text, they preprocess transcribed video narrations by discarding common stop-words and utilizing the GoogleNews pretrained word2vec embedding model for word representations. The joint embedding model maps these video and text features into a common dimensional space using non-linear embedding functions, including a linear fully connected layer and a context gating function. Training is guided by a max-margin ranking loss, ensuring higher similarity for matching video-caption pairs, with an intra-video negative sampling strategy to emphasize relevant video aspects over background features.

In the realm of surgery, Yuan et al. (2023) developed the Surgical Vision Language Pre-training (SurgVLP), a

surgical vision-language foundation model pre-trained on surgical video lectures. With a CLIP-like architecture, the Surgvlp model uses a ResNet50 as an image encoder and a transformer-based text encoder. For the pre-training, they utilize the Surgical Video Lecture (SVL) dataset, which consists of surgical video lectures sourced from e-learning platforms with captions automatically generated with ASR systems. Specifically, the AWS medical transcribe ASR system is used for medical terminology and surgery-specific terms, and the Whisper ASR system is used for general sentence structure and common words since the AWS ASR generates incomplete sentence fragments. Like the CLIP pretraining, the objective focuses on the cosine similarity between the video-text pairs by employing the InfoNCE loss (Oord et al., 2018) combined with the MIL-NCE loss (Miech et al., 2020).

In Yuan et al. (2024), a novel approach called HecVL (Yuan et al., 2024) was introduced, which leverages hierarchical video-language pretraining to build a generalist surgical model. This model uses a hierarchical video-text paired dataset, pairing surgical lecture videos with three hierarchical levels of texts: clip-level transcribed audio texts for atomic actions, phase-level conceptual text summaries, and video-level abstract text of the surgical procedure. The HecVL employs a fine-to-coarse contrastive learning framework to learn separate embedding spaces for these hierarchies within a single model. This disentangling of embedding spaces allows the model to encode short-term and long-term surgical concepts. Injecting textual semantics enables HecVL to perform zero-shot surgical phase recognition without any human annotation, and it demonstrates the ability to transfer the same model across different surgical procedures and medical centers, showcasing its robustness and versatility.

### *2.3. Prompt Tuning*

Large pre-trained vision-language models like CLIP can transfer learned representations across diverse downstream tasks. These models align images and texts within a shared feature space, facilitating zero-shot transfer via prompting. However, prompt engineering is a major challenge in zero-shot transfer, which is time-consuming and requires substantial domain expertise due to performance sensitivity to wording changes.

Context Optimization (CoOp) proposed by Zhou et al. (2022c) is a significant advancement addressing the prompt engineering challenge by introducing learnable vectors for the prompt's context words while keeping the pre-trained parameters fixed. CoOp's two implementations—unified context and class-specific context—have demonstrated substantial improvements over hand-crafted prompts with minimal labeled data, achieving notable performance gains and excellent domain generalization across 11 datasets. However, CoOp's static prompts can overfit base classes, reducing generalizability to unseen classes within the same dataset. Conditional Context Optimization (CoCoOp) by Zhou et al. (2022a) extends CoOp by incorporating a lightweight neural network that generates an input-conditional token for each image, enabling dynamic prompts that adapt to each instance. This approach mitigates the overfitting issue seen in CoOp, significantly improving generalization to unseen classes and demonstrating strong transferability beyond single datasets.

Knowledge-aware Prompt-tuning (KnowPrompt) proposed by Chen et al. (2022) and Knowledge-Aware Prompt Tuning (KAPT) by Kan et al. (2023) further advance prompt learning by integrating external knowledge. KnowPrompt injects latent knowledge from relation labels into prompt construction, using learnable virtual type words and answer words optimized with structured constraints. This approach has shown effectiveness in relation to extraction tasks across multiple datasets, particularly in low-resource settings. KAPT, inspired by human intelligence, uses both discrete and continuous prompts to leverage external knowledge, enhancing few-shot image classification and improving generalization to unseen categories. KAPT outperforms state-of-the-art methods like CoCoOp, achieving significant gains in new class recognition.

K-LITE (Knowledge-augmented Language-Image Training and Evaluation) Shen et al. (2022) proposes a strategy to incorporate external structured knowledge, such as WordNet and Wiktionary, into the training and evaluation of vision-language models. This method enhances entity descriptions with additional knowledge, improving image representation learning and facilitating zero-shot and few-shot transfers. K-LITE has considerably improved image classification and object detection across numerous datasets.

In the medical domain, leveraging external knowledge through well-designed prompts has proven crucial for transferring knowledge from pre-trained vision-language models. Studies show that medical prompts, enriched with expert-level knowledge and image-specific information, significantly improve zero-shot performance. Methods for the automatic generation of medical prompts can further enhance fine-grained grounding, demonstrating the broad applicability of these approaches across various medical imaging modalities. Qin et al. (2022) proposes a combination of a VQA model with the PubMedBert transformer to generate prompts based on the physical attributes of polyps, mimicking the description of medical professionals.

These advancements in prompt tuning, prompt learning, and integrating external knowledge highlight the evolution of vision-language models. By addressing the limitations of manual prompt engineering, enhancing model adaptability and generalization, and improving efficiency and effectiveness across diverse domains, these techniques offer promising alternatives to tradi-

tional fine-tuning when adapting vision-language foundation models to downstream tasks.

## 3. Material and methods

This section presents the proposed approach for adapting a vision-language model for surgical phase recognition.

### *3.1. Data*

#### *3.1.1. Image Dataset*

The dataset used in this project is the Cholec80 dataset, which contains 80 videos of laparoscopic cholecystectomy surgery divided into seven surgical phases. The videos were captured at 25 frames per second, with a resolution of $1920 \times 1080$. The videos last an average of 38 minutes with a 16-minute standard deviation. Table 1 lists each of the seven surgical phases and their average duration across all videos. The frames were extracted for processing, totaling 1000 frames per video. From the 80 videos, 40 were used for training, 8 for validation, and 32 for testing.

| Phase | Duration (s) |
|---|---|
| Preparation | 125±95 |
| Calot triangle dissection | 954±538 |
| Clipping and cutting | 168±152 |
| Gallbladder dissection | 857±551 |
| Gallbladder packaging | 98±53 |
| Cleaning and coagulation | 178±166 |
| Gallbladder retraction | 83±56 |

Table 1: List of each surgical phase in the Cholec80 dataset with the mean and standard deviation of duration

#### *3.1.2. Text Prompts*

A hand-crafted textual prompt describing each class label in a single sentence was created for image pairings. These prompts were manually generated to expand the class label by including the associated surgical tools and critical attributes, ensuring each class is unique and distinguishable. Additionally, knowledge-infused prompts were generated using large language models (LLMs), specifically GPT-4 and GPT-4o (Achiam et al., 2023). The hand-crafted prompts, along with a medical textbook detailing the procedure, were fed into the LLMs to produce new prompts. This approach simulates how a professional would utilize their innate knowledge to craft and refine prompts, leveraging the comprehensive information from the textbook and the extensive pre-trained data in the LLMs. While the hand-crafted prompt is composed of one sentence with an average of 16 words, the LLM-generated prompts are considerably bigger, with the GPT-4 and GPT-4o prompts composed of a minimum of two sentences with 43 words and 65 words average, respectively.

### *3.2. Vision-Language Model Architecture*

The pre-trained model selected is the SurgVLP proposed in Yuan et al. (2023). Inspired by the CLIP architecture (Radford et al., 2021), it comprises visual and text encoders. The selected visual encoder is the ResNet-50 model that employs a stack of 50 layers with residual learning, utilizing skip connections (He et al., 2016), pre-trained on the ImageNet dataset as a base for the pre-training. The text encoder is the BioClinicalBert (Huang et al., 2019), a base-size Bert model containing 12 encoders with 12 bidirectional self-attention heads totaling 110 million parameters. It is pre-trained on the MIMIC-III dataset (Johnson et al., 2016), which consists of 2083180 clinical notes.

The Surgical Video Lecture (SVL) dataset (Yuan et al., 2023) is utilized during pre-training and comprises video-text pairs of surgical video lectures their corresponding transcriptions. During training, the embedding spaces between the visual encoder and text encoder are aligned by jointly training them to maximize the cosine similarity of the video and text embeddings between the real video-text pairs while minimizing the cosine similarity between the incorrect pairings. It employs the InfoNCE (Oord et al., 2018) loss function, a type of contrastive loss commonly used in self-supervised learning and multi-modal representation to align the video-text pairs combined with MIL-NCE (Miech et al., 2020) learning objectives to address the misalignment issue that the lecturers might talk about previously or after the visual demonstration.

### *3.3. Vision-Language Model Adaptation to Downstream task*

A specialized adaptation head is necessary to adapt a foundation model for a downstream task. This head, as depicted in Figure 3, uses a feature fusion technique to combine the image and text outputs from the backbones, tailoring them to the specific task, such as phase recognition. This project investigates two feature fusion techniques: the weighted sum of features and a gated mechanism. Subsequently, the visual encoder is fine-tuned alongside the adaptation head.

One strategy is the weighted sum method, illustrated in Figure 4, which computes the similarity between image and text features, applies weighted text features to the image features and combines them for classification outputs. First, the image $I_{2048}$ and text $T_{7\times2048}$ feature vectors are normalized. The normalized text features are then transposed along the last two dimensions to obtain $T_{2048\times7}$. To facilitate matrix multiplication, an extra dimension is added to the image feature vector, resulting in $I_{1\times2048}$. Next, the cosine similarity between

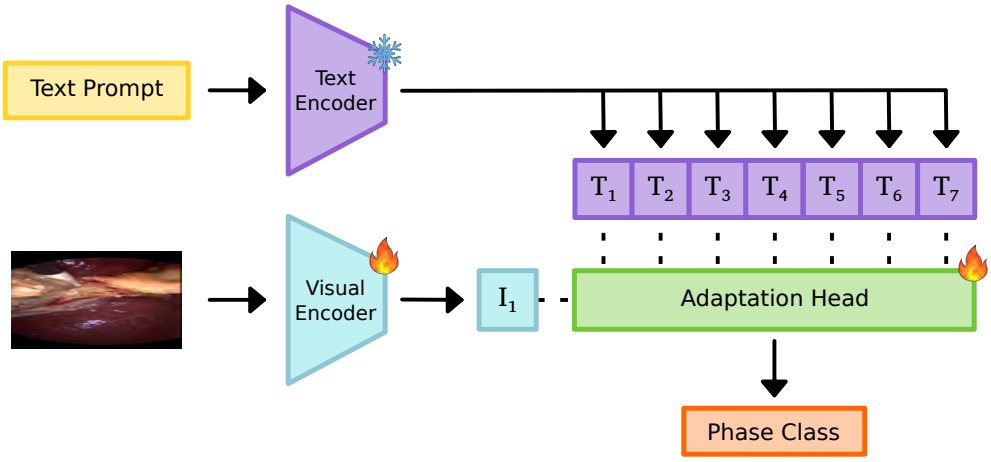



Figure 3: SurgVLP model architecture with adaptation module

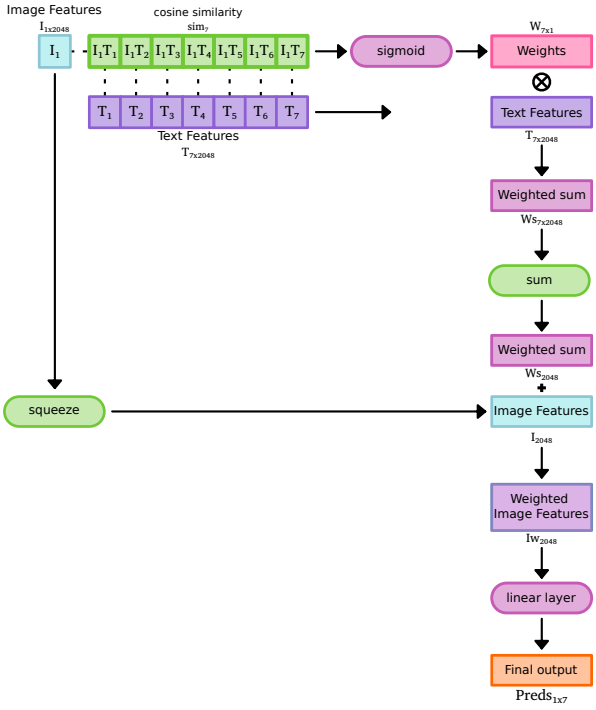



Figure 4: Weighted sum feature aggregation method

the image features and each text feature is calculated, yielding $sim_7$. A softmax function is applied to these similarities, and the resulting values are expanded along the last dimension to generate weights $W_{7\times1}$. These weights are then multiplied by the text features to produce the weighted text features $Tw_{7\times2048}$, which are summed along axis 1 to derive the final weighted text features $Tw_{2048}$. Finally, the weighted text features are summed with the image features $I_{2048}$ to obtain the combined image features $Iw_{2048}$. A linear layer is then applied to these combined features for phase classification, resulting in the predictions $preds_7$.

The other strategy is the gated mechanism, illustrated in Figure 5, which employs gates to dynamically control the contribution of image and text features to the final representation. Initially, to obtain the gate weights, the image feature vector $I_{1\times2048}$ and the text feature vector $T_{7\times2048}$ are concatenated on their last axis, forming the combined feature vector $C_{7\times4096}$. A linear layer is then applied to reduce the dimensionality to $C_{7\times2048}C$, followed by a sigmoid function to generate the gate weights $W_{7\times2048}$. These weights are then multiplied by the image feature vectors $W \times I$, resulting in the gated output 1 $G1_{7\times2048}$. Simultaneously, the complementary weights $1-W$ are multiplied by the text feature vector $(1-W)\times T$, producing the gated output 2 $G2_{7\times2048}$. The final gated representation $G_{7\times2048}$ is obtained by summing both outputs: $G = G1 + G2$. Subsequently, mean pooling is performed along axis 1 to average across each textual feature representation, resulting in the final output for classification $G_{2048}$. A linear layer is then applied to these combined features for phase classification, culminating in the predictions $preds_7$.

### 3.4. Proposed Experiments

This section outlines the experiments conducted to investigate the impact of the text prompts on the adaptation of the SurgVLP model for the surgical phase recognition task.

#### 3.4.1. Textual and visual feature's dimensionality

Phase recognition is a task that relies heavily on video input; it is necessary to preserve as much of the visual input as possible. Due to the dimensionality output of the encoders in each branch, image encoder $I_{2048}$ and text encoder $T_{7\times768}$, one of the output shapes must be manipulated to match the other. One of the experiments performed is to analyze the advantages of up-scaling the text feature vector to match the image feature vector dimensions instead of down-scaling the image features to the text vector. A linear layer was applied to downscale the image vector $I_{2048}$ to $I_{768}$ or upscale the text vector $T_{7\times768}$ to $T_{7\times2048}$.

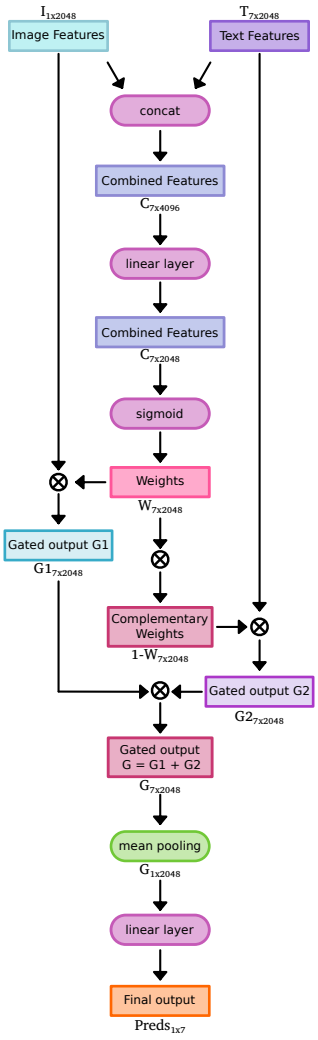



Figure 5: Gated mechanism feature aggregation method

### 3.4.2. *Adding More Textual Information*

In addition to generating prompts with an LLM, we explore methods to enhance hand-crafted prompts with additional information. Specifically, detailed descriptions of each class label were extracted from a medical textbook (Majumder et al., 2020) and processed through a text encoder. These textual features from the textbook were then combined with the hand-crafted text features using an attention mechanism, as illustrated in Figure 6. First, the hand-crafted prompt passes through the SurgVLP text encoder to obtain text features $Ta_{7\times768}$; in parallel, the textbook text passes through another text encoder to obtain text features $Tb_{7\times768}$. Similarly to the previously described gate mechanism, both text features are concatenated at their last dimension, obtaining the combined feature vector $C_{7\times1536}$. A linear layer is applied to reduce the dimensionality to $C_{7\times1}$ followed by a sigmoid function to generate the attention weights $W_{7\times1}$. These weights are then expanded to match the text feature dimensionality $W_{7x768}$ and multiplied by the $Ta_{7x768}$ obtaining the attended output $A1_{7x768} = W_{7x768} \times Ta_{7x768}$. Simultaneously, the complementary weights $1 - W$ are multiplied by $Tb_{7x768}$ obtaining the attended output $A2_{7x768} = (1 - W_{7x768}) \times Tb_{7x768}$. Both outputs are then summed to obtain the final attended text features $A_{7x768} = A1_{7x768} + A2_{7x768}$, which are then used as the text features to combine with the image features.

## 4. Results

Experiments were conducted on 100%, 25%, and 12.5% of the training data to analyze trends across the evaluated techniques. For the 25% and 12.5% splits, three folds of randomly selected videos from the training set were chosen, ensuring no overlap among them. The evaluation metrics observed are accuracy, f1-score, precision, and recall described in Equations 1 - 4, where $tp, tn, fp, fn$ represent *true positive*, *true negative*, *false positive* and *false negative* respectively. These metrics were calculated per video in the test set, and the results were then averaged to provide a comprehensive assessment.

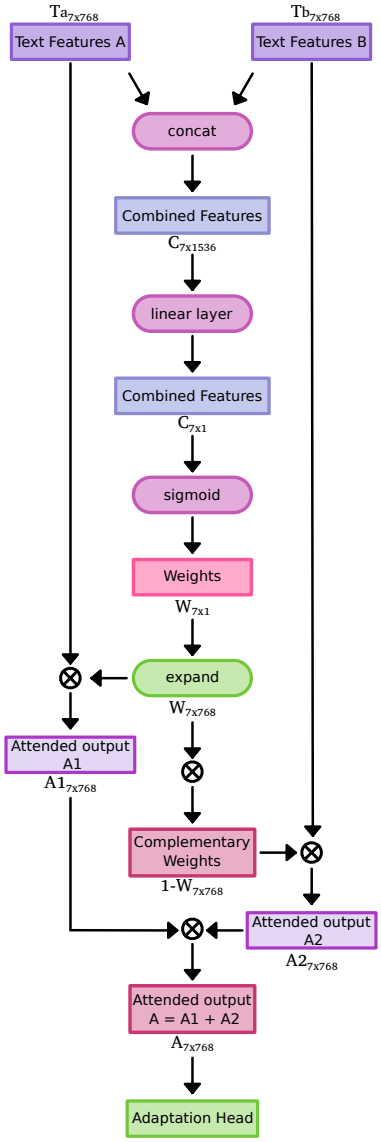



Figure 6: Attention mechanism feature combination technique

All experiments were done using Pytorch as the main deep learning framework and run on Nvidia V100 and RTX6000 GPUs provided by the University of Strasbourg.

$$Acuracy = \frac{tp + tn}{tn + tp + fn + fp} \quad (1)$$

$$F1 - score = 2 \cdot \left( \frac{Precision \cdot Recall}{Precision + Recall} \right) \quad (2)$$

$$Recall = \frac{tp}{tp + fn} \quad (3)$$

$$Precision = \frac{tp}{tp + fp} \quad (4)$$

| Parameters | values |
|---|---|
| *Input dim.* | 224 x 224 |
| *Batch size* | 64 |
| *Optimizer* | *SGD* |
| *Learning rate* | 0.001 |
| *LR scheduler* | *Cosine Annealing* |
| *Loss function* | *Cross Entropy* |

Table 2: Final hyperparameters selected

### *4.1. Vision vs Vision Language*

To evaluate the efficacy of the SurgVLP model pre-trained on the SVL dataset, we fine-tuned the SurgVLP model's visual encoder and a ResNet50 pre-trained on the IMAGENET dataset for phase recognition. Var-

ious hyperparameters were explored, including batch sizes of 32 and 64, optimizers Adam and SGD, using a cosine annealing learning rate scheduler, and learning rates ranging from 0.0001 to 0.01. Table 2 details the final selected hyperparameters. The SurgVLP's model was also fine-tuned, integrating the text features from the frozen text encoder as depicted in Figure 3 using the two previously described fusion techniques. The final results can be observed in Table 3.

The domain-specific pre-training of SurgVLP (vision only) results in a slight improvement over Resnet-50, with higher accuracy (78.24% vs. 77.99%) and recall (78.60% vs. 77.45%), when trained on 100% of the data indicating that pre-training on surgical data enhances the model's ability to adapt to surgical contexts. When integrating textual data, the full SurgVLP model using a weighted sum aggregation of image-text features outperforms the vision-only and Resnet-50 models, achieving the highest accuracy (78.68%) and F1-score (72.23%). Although slightly less effective than the weighted sum, the gated mechanism for image-text feature aggregation in SurgVLP still shows competitive performance.

In few-shot training scenarios, the performance gaps become more pronounced. With only 25% of the training data, SurgVLP (vision only) maintains a significant edge over Resnet-50, showing a higher accuracy (66.52% vs. 61.61%) and recall (69.26% vs. 63.45%). The full SurgVLP model with weighted sum aggregation further extends this lead (67.84% accuracy), demonstrating the added value of incorporating textual features. However, the gated mechanism's performance drops notably in the 12.5% training split, indicating potential instability with minimal data. Here, the weighted sum aggregation maintains robust performance (58.43% accuracy), whereas Resnet-50's accuracy falls to 56.62%. These observations underscore the importance of domain-specific pre-training and effective image-text feature aggregation, particularly in data-constrained environments.

### *4.2. Textual and visual feature's dimensionality*

To explore the impact on the feature dimensionality when combining image and text features, the model's visual encoder and adaptation head with the weighted sum fusion technique were fine-tuned following the same parametrization as depicted in 4.1. The final results are exemplified in Table 4.

With the entire training dataset, the model using upscaled text features marginally outperforms the downscaled image feature model, achieving higher accuracy (78.68% vs. 78.34%) and F1-score (72.23% vs. 71.74%). However, as the training data decreases, the performance gap widens significantly. With only 25% of the training data, the upscaled text feature model retains a higher accuracy (67.84% vs. 66.39%) and F1-score (64.45% vs. 63.10%). This trend is even more pronounced in the 12.5% training scenario, where the upscaled text feature model maintains a reasonable accuracy (58.43%) and F1-score (58.60%). In contrast, the downscaled image feature model's accuracy drops to 46.94% with a notable decrease in the F1-score (52.67%). These results suggest that maintaining higher dimensionality in textual features enhances the model's robustness, particularly in low-data settings, likely due to better preservation of relevant visual features critical for accurate surgical phase recognition.

### *4.3. Text Prompt Generation*

This experiment investigates the capabilities of a large language model to generate static prompts by simulating the expert prompt-making process. We fine-tuned the image backbone and the adaptation head using the weighted sum fusion technique for the three discussed prompts. Figure 7 highlights the cosine similarity between classes in each text prompt to evaluate class differentiability and identify potential misrepresentation. The results are displayed in Table 5.

When trained on the entire dataset, the differences between hand-crafted and GPT-generated prompts are minimal, with the GPT-generated prompts slightly edging out in accuracy and F1-score. Specifically, the GPT4o prompt achieved an accuracy of 78.70% and an F1-score of 72.25%, compared to 78.68% and 72.23% for the hand-crafted prompt. However, as the training data decreases, the impact of the prompt choice becomes more apparent. With 25% of the data, the GPT4o prompt maintains a slight advantage, showing a higher accuracy (67.93% vs. 67.84%) and F1-score (64.49% vs. 64.45%). In the most data-constrained scenario (12.5% of the data), all prompts perform similarly. However, the GPT4o prompt continues to show better results in terms of consistency and robustness, with an F1-score of 58.61%. These results suggest that while hand-crafted and GPT-generated prompts have a negligible impact with abundant data, the GPT4o prompt provides a slight performance edge in few-shot learning scenarios.

### *4.4. Adding more text*

This experiment evaluated two text encoders: the SurgVLP pre-trained text encoder, which has its embedding space aligned with the image encoder, and the ClinicalBert transformer pre-trained on the MIMIC-III dataset. The textual features were extracted from the textbook prompt and combined with the text features from the hand-crafted prompts through the attention mechanism previously described. The attended features were combined with the image features utilizing the weighted sum fusion technique and fine-tuned with the image encoder and it's results are depicted in Table 6.

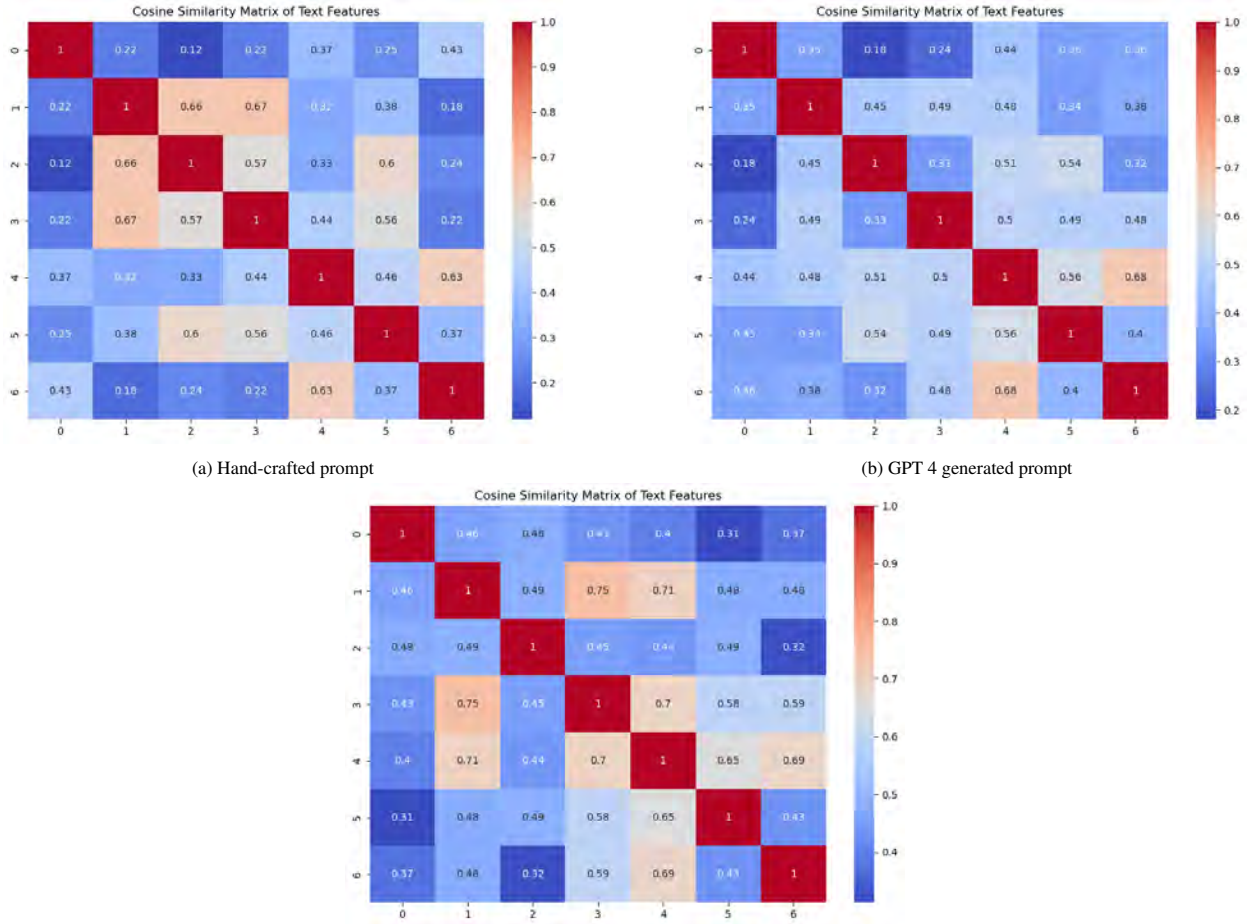
As shown in Table 6, when trained on the full dataset (100%), the addition of textbook knowledge

| Model | Training Split | Dataset | Accuracy(%) | F1-score(%) | Precision(%) | Recall(%) |
|---|---|---|---|---|---|---|
| *Resnet50* | 100% | IMAGENET | 77.99 ± 9 | 71.41 ± 7 | 66.59 ± 9 | 77.45 ± 6 |
| *SurgVLP (vision only)* | 100% | SVL | 78.24 ± 9 | 71.66 ± 7 | 66.27 ± 9 | 78.60 ± 6 |
| *SurgVLP (weighted sum)* | 100% | SVL | **78.68 ± 9** | **72.23 ± 7** | **67.00 ± 9** | **78.93 ± 6** |
| *SurgVLP (gated mechanism)* | 100% | SVL | 78.02 ± 9 | 71.61 ± 7 | 66.15 ± 9 | 78.63 ± 6 |
| *Resnet50* | 25% | IMAGENET | 61.61 ± 5 | 58.30 ± 3 | 54.40 ± 3 | 63.45 ± 4 |
| *SurgVLP (vision only)* | 25% | SVL | 66.52 ± 1 | 63.82 ± 1 | 59.55 ± 1 | 69.26 ± 3 |
| *SurgVLP (weighted sum)* | 25% | SVL | **67.84 ± 0.2** | **64.45 ± 1** | **59.79 ± 1** | **70.37 ± 2** |
| *SurgVLP (gated mechanism)* | 25% | SVL | 62.07 ± 3 | 61.36 ± 2 | 56.10 ± 1 | 68.23 ± 3 |
| *Resnet50* | 12.5% | IMAGENET | 56.62 ± 0.4 | 53.91 ± 2 | 51.00 ± 1 | 57.73 ± 2 |
| *SurgVLP (vision only)* | 12.5% | SVL | 57.77 ± 2 | 58.17 ± 1 | 54.03 ± 0.2 | 63.53 ± 1 |
| *SurgVLP (weighted sum)* | 12.5% | SVL | **58.43 ± 2** | **58.60 ± 1** | **54.44 ± 1** | **64.01 ± 2** |
| *SurgVLP (gated mechanism)* | 12.5% | SVL | 28.54 ± 14 | 38.56 ± 14 | 36.60 ± 13 | 42.42 ± 15 |

Table 3: Fine-tuning results for different training splits and feature dimensions

| Model | Training Split | Feature Dimension | Accuracy(%) | F1-score(%) | Precision(%) | Recall(%) |
|---|---|---|---|---|---|---|
| *SurgVLP (Upscale text)* | 100% | $T_{7x2048}$ | **78.68 ± 9** | **72.23 ± 7** | **67.00 ± 9** | **78.93 ± 6** |
| *SurgVLP (Downscale image)* | 100% | $I_{1x768}$ | 78.34 ± 9 | 71.74 ± 7 | 66.70 ± 9 | 78.17 ± 7 |
| *SurgVLP (Upscale text)* | 25% | $T_{7x2048}$ | **67.84 ± 0.2** | **64.45 ± 1** | **59.79 ± 1** | **70.37 ± 2** |
| *SurgVLP (Downscale image)* | 25% | $I_{1x768}$ | 66.39 ± 1 | 63.10 ± 1 | 58.32 ± 1 | 69.23 ± 2 |
| *SurgVLP (Upscale text)* | 12.5% | $T_{7x2048}$ | **58.43 ± 2** | **58.60 ± 1** | **54.44 ± 1** | **64.01 ± 2** |
| *SurgVLP (Downscale image)* | 12.5% | $I_{1x768}$ | 46.94 ± 4 | 52.67 ± 3 | 49.45 ± 3 | 56.87 ± 3 |

Table 4: Results exploring different feature dimensions and training splits




(a) Hand-crafted prompt

(b) GPT 4 generated prompt

(c) GPT 4o generated prompt

Figure 7: Cosine similarity between surgical phase classes for each text prompts

| Text Prompt | Training Split | Dataset | Accuracy(%) | F1-score(%) | Precision(%) | Recall(%) |
|---|---|---|---|---|---|---|
| *hand-crafted* | 100% | SVL | 78.68 ± 9 | 72.23 ± 7 | 67.00 ± 9 | 78.93 ± 6 |
| *gpt4* | 100% | SVL | 78.70 ± 9 | 72.22 ± 7 | 67.00 ± 9 | 78.89 ± 6 |
| *gpt4o* | 100% | SVL | 78.70 ± 9 | 72.25 ± 7 | 67.00 ± 9 | 78.95 ± 6 |
| *hand-crafted* | 25% | SVL | 67.84 ± 0.2 | 64.45 ± 1 | 59.79 ± 1 | 70.37 ± 2 |
| *gpt4* | 25% | SVL | 67.83 ± 0.2 | 64.47 ± 1 | 59.80 ± 1 | 70.39 ± 2 |
| *gpt4o* | 25% | SVL | 67.93 ± 0.2 | 64.49 ± 1 | 59.73 ± 1 | 70.54 ± 2 |
| *hand-crafted* | 12.5% | SVL | 58.43 ± 2 | 58.60 ± 1 | 54.44 ± 1 | 64.01 ± 2 |
| *gpt4* | 12.5% | SVL | 58.45 ± 2 | 58.60 ± 1 | 54.44 ± 1 | 64.01 ± 2 |
| *gpt4o* | 12.5% | SVL | 58.44 ± 2 | 58.61 ± 1 | 54.44 ± 1 | 64.01 ± 2 |

Table 5: Results with different text prompts for various training splits

slightly decreased performance across both encoders, with SurgVLP maintaining a marginally higher accuracy (78.59%) and F1-score (72.24%) compared to ClinicalBERT (78.53% accuracy and 72.16% F1-score). However, these differences become more significant in few-shot learning scenarios.

With 25% of the training data, the SurgVLP text encoder's performance remained stable (67.82% accuracy and 64.45% F1-score), closely matching the baseline handcrafted prompt performance (67.84% accuracy). In contrast, the ClinicalBERT's performance dropped sharply to 58.44% accuracy and 58.61% F1-score, demonstrating its less effective adaptation to surgical phase recognition when compared to SurgVLP.

The performance gap is even more pronounced, with only 12.5% of the training data. The SurgVLP

text encoder showed a notable increase in accuracy (61.61%) and maintained a competitive F1-score (58.30%), whereas the ClinicalBERT model's accuracy and F1-score (both 58.44% and 58.61%) mirrored the results of using no additional text. These results indicate that the SurgVLP text encoder, explicitly trained on surgical lecture transcriptions, better preserves relevant surgical context and adapts more effectively in low-data environments than ClinicalBERT, underscoring the importance of domain-specific pre-training for surgical phase recognition.

## 5. Discussion

The experiments conducted in this study demonstrated the effectiveness of adapting generalist vision-language models for surgical phase recognition tasks. The SurgVLP model, pre-trained on a comprehensive dataset of surgical video lectures, showed promising results in identifying different phases of laparoscopic cholecystectomy. The model achieved high accuracy and robust performance across various training data percentages by integrating text prompts and employing feature fusion techniques.

### *5.1. Vision-Language Models vs. Vision-Only Models*

One of the significant findings from our experiments is the comparative performance of vision-language and vision-only models. The SurgVLP model, which incorporates both visual and textual information, consistently outperformed the ResNet50 vision-only model. The results indicate that including text prompts provides additional context that enhances the model's ability to distinguish between different surgical phases. Specifically, the SurgVLP model with the weighted sum feature fusion technique achieved the highest accuracy and F1-score across all training splits.

### *5.2. Impact of Feature Fusion Techniques*

The study explored two feature fusion techniques: weighted sum and gated mechanisms. The weighted sum method demonstrated superior performance compared to the gated mechanism. This finding suggests that the direct combination of text and image features, weighted by their similarity, is more effective in capturing the relevant information needed for phase recognition. The gated mechanism, while useful, may introduce unnecessary complexity that does not translate into improved performance for this specific task.

### *5.3. Upscaling vs. Downscaling Features*

The study also compared the effects of upscaling text features to match the image feature dimensions versus downscaling image features to match text dimensions. The results favored upscaling text features, which maintained higher accuracy and F1-scores. This outcome underscores the importance of preserving as much visual information as possible, given that surgical phase recognition relies heavily on video inputs.

### *5.4. Text Prompt Generationg*

Another critical aspect investigated was the generation of text prompts using large language models (LLMs). Hand-crafted prompts, GPT-4 generated prompts, and GPT-4o generated prompts were compared. Analyzing the cosine similarity between the classes for each prompt, depicted in Figure 7, shows the differentiability between each class. Although GPT-4 prompts show a more considerable differentiation between each class, it does not correlate to a better prompt as shown in Tables 5 where no significant discrepancy can be observed. Interestingly, the performance differences among these prompts were minimal, indicating that even simple, well-constructed prompts can be as effective as those generated by advanced LLMs. This finding is significant as it suggests that substantial domain expertise may not be necessary to generate effective text prompts, potentially lowering the barrier to implementing such systems in clinical settings. Also, there is no significant impact on the number of tokens as input for the text encoder since the LLM-generated prompts are considerably more verbose than the hand-crafted prompt.

### *5.5. Adding More Textual Information*

Incorporating additional textual information from medical textbooks into the text prompts did not significantly enhance the model's performance. This result suggests that the initial prompts were already sufficiently informative and that adding more text did not provide additional benefits. It also highlights the robustness of the initial prompt design and the efficiency of using concise, targeted information.

## 6. Conclusions

This research focuses on adapting generalist vision-language models, specifically the SurgVLP model, for surgical phase recognition in laparoscopic cholecystectomy. The study aims to enhance surgical safety, efficiency, and outcomes by developing AI systems that can be easily adapted to various surgical tasks. Key objectives include exploring feature fusion techniques, evaluating text prompt generation methods, assessing the model's performance with different training data amounts, and integrating additional textual information from medical literature.

The experiments demonstrated that the SurgVLP model, which incorporates both visual and textual information, consistently outperformed the ResNet50 vision-only model. The weighted sum feature fusion technique

| Additional Prompt | Training Split | Dataset | Accuracy(%) | F1-score(%) | Precision(%) | Recall(%) |
|---|---|---|---|---|---|---|
| - | 100% | - | **78.68 ± 9** | **72.23 ± 7** | **67.00 ± 9** | **78.93 ± 6** |
| *textbook* | 100% | SurgVLP | 78.59 ± 9 | 72.24 ± 7 | 67.00 ± 9 | 78.95 ± 6 |
| *textbook* | 100% | ClinicalBert | 78.53 ± 9 | 72.16 ± 7 | 66.92 ± 9 | 78.85 ± 6 |
| - | 25% | - | **67.84 ± 0.2** | **64.45 ± 1** | **59.79 ± 1** | **70.37 ± 2** |
| *textbook* | 25% | SurgVLP | 67.82 ± 0.2 | 64.45 ± 1 | 59.78 ± 1 | 70.37 ± 2 |
| *textbook* | 25% | ClinicalBert | 58.44 ± 2 | 58.61 ± 1 | 54.44 ± 1 | 64.01 ± 2 |
| - | 12.5% | - | 58.43 ± 2 | 58.60 ± 1 | **54.44 ± 1** | **64.01 ± 2** |
| *textbook* | 12.5% | SurgVLP | **61.61 ± 5** | 58.30 ± 3 | 54.40 ± 3 | 63.45 ± 4 |
| *textbook* | 12.5% | ClinicalBert | 58.44 ± 2 | **58.61** ± 1 | **54.44 ± 1** | **64.01 ± 2** |

Table 6: Results when adding more information for 100% of the training videos

showed superior performance to the gated mechanism, suggesting that the direct combination of text and image features is more effective for phase recognition. Text prompts generated by large language models (LLMs) like GPT-4 and GPT-4o did not significantly outperform hand-crafted prompts, indicating that simple, well-constructed prompts can be just as effective. Additionally, upscaling text features to match image feature dimensions maintained higher accuracy and F1-score, underscoring the importance of preserving visual information.

While the results are promising, this study has limitations. The experiments were conducted using a specific surgical procedure (laparoscopic cholecystectomy), and the generalizability to other types of surgeries needs further investigation. Additionally, this study adapts the vision-language model by extracting features from surgical videos on a frame level, losing any temporal information from the procedure.

For future research, this study could be enhanced by including a broader range of surgical procedures to increase the comprehensiveness of the analysis. Exploring methods to integrate temporal information into the pipeline could also provide further improvements.

## 7. Acknowledgments

This work has received funding from the European Union (ERC, CompSURG, 101088553). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This work was also partially supported by French state funds managed by the ANR under Grant ANR-10-IAHU-02. It was granted access to the HPC resources of Unistra Mesocentre.

I want to thank my supervisors, Nicolas, Vinkle, and Kun, for all their time, attention, encouragement, and patience during my master's thesis project. Thank you to the MAIA program for granting me this fantastic opportunity to study and learn from brilliant minds and rich cultures. Thank you to all my friends worldwide for always being a source of inspiration and courage, supporting me during the tough times, and laughing alongside me during the easier ones. Para meus pais, irmãos, tios e primos cujo apoio e amor me permitem seguir este sonho mesmo em frente a uma saudade imensurável, muito obrigada.

## References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al., 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774 .

Buskens, C.J., Sahami, S., Tanis, P.J., Bemelman, W.A., 2014. The potential benefits and disadvantages of laparoscopic surgery for ulcerative colitis: a review of current evidence. Best Practice & Research Clinical Gastroenterology 28, 19–27.

Chen, X., Zhang, N., Xie, X., Deng, S., Yao, Y., Tan, C., Huang, F., Si, L., Chen, H., 2022. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction, in: Proceedings of the ACM Web conference 2022, pp. 2778–2788.

Coccolini, F., Catena, F., Pisano, M., Gheza, F., Fagiuoli, S., Di Saverio, S., Leandro, G., Montori, G., Ceresoli, M., Corbella, D., et al., 2015. Open versus laparoscopic cholecystectomy in acute cholecystitis. systematic review and meta-analysis. International journal of surgery 18, 196–204.

Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S.T., Navab, N., 2020. Tecno: Surgical phase recognition with multi-stage temporal convolutional networks, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23, Springer. pp. 343–352.

Eisenmann, M., Reinke, A., Weru, V., Tizabi, M.D., Isensee, F., Adler, T.J., Godau, P., Cheplygina, V., Kozubek, M., Ali, S., et al., 2022. Biomedical image analysis competitions: The state of current participation practice. arXiv preprint arXiv:2212.08568 .

Golany, T., Aides, A., Freedman, D., Rabani, N., Liu, Y., Rivlin, E., Corrado, G.S., Matias, Y., Khoury, W., Kashtan, H., et al., 2022. Artificial intelligence for phase recognition in complex laparoscopic cholecystectomy. Surgical Endoscopy 36, 9215–9223.

Han, H.S., Shehta, A., Ahn, S., Yoon, Y.S., Cho, J.Y., Choi, Y., 2015. Laparoscopic versus open liver resection for hepatocellular carcinoma: case-matched study with propensity score matching. Journal of hepatology 63, 643–650.

Hashimoto, D.A., Rosman, G., Rus, D., Meireles, O.R., 2018. Artifi-

cial intelligence in surgery: promises and perils. Annals of surgery 268, 70–76.

Hassler, K., Collins, J., Philip, K., et al., 2021. Laparoscopic cholecystectomy.[updated 2023 jan 23]. StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing .

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.

Huang, K., Altosaar, J., Ranganath, R., 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342 .

Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T., 2021. Scaling up visual and vision-language representation learning with noisy text supervision, in: International conference on machine learning, PMLR. pp. 4904–4916.

Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G., 2016. Mimic-iii, a freely accessible critical care database. Scientific data 3, 1–9.

Kan, B., Wang, T., Lu, W., Zhen, X., Guan, W., Zheng, F., 2023. Knowledge-aware prompt tuning for generalizable vision-language models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15670–15680.

Kitaguchi, D., Takeshita, N., Matsuzaki, H., Takano, H., Owada, Y., Enomoto, T., Oda, T., Miura, H., Yamanashi, T., Watanabe, M., et al., 2020. Real-time automatic surgical phase recognition in laparoscopic sigmoidectomy using the convolutional neural network-based deep learning approach. Surgical endoscopy 34, 4924–4931.

Lam, K., Qiu, J., 2024. Foundation models: the future of surgical artificial intelligence? British Journal of Surgery 111, znae090.

Majumder, A., Altieri, M.S., Brunt, L.M., 2020. How do i do it: laparoscopic cholecystectomy. Annals of Laparoscopic and Endoscopic Surgery 5.

Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A., 2020. End-to-end learning of visual representations from uncurated instructional videos, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9879–9889.

Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J., 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips, in: ICCV.

Olsen, D.O., 1991. Laparoscopic cholecystectomy. The American journal of surgery 161, 339–344.

Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 .

Padoy, N., Blum, T., Ahmadi, S.A., Feussner, H., Berger, M.O., Navab, N., 2012. Statistical modeling and recognition of surgical workflow. Medical image analysis 16, 632–641.

Qin, Z., Yi, H., Lao, Q., Li, K., 2022. Medical image understanding with pretrained vision language models: A comprehensive study. arXiv preprint arXiv:2209.15517 .

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR. pp. 8748–8763.

Shabanzadeh, D.M., Christensen, D.W., Ewertsen, C., Friis-Andersen, H., Helgstrand, F., Nannestad Jørgensen, L., Kirkegaard-Klitbo, A., Larsen, A.C., Ljungdalh, J.S., Nordblad Schmidt, P., et al., 2022. National clinical practice guidelines for the treatment of symptomatic gallstone disease: 2021 recommendations from the danish surgical society. Scandinavian Journal of Surgery 111, 11–30.

Shen, S., Li, C., Hu, X., Xie, Y., Yang, J., Zhang, P., Gan, Z., Wang, L., Yuan, L., Liu, C., et al., 2022. K-lite: Learning transferable visual models with external knowledge. Advances in Neural Information Processing Systems 35, 15558–15573.

Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N., 2016. Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging 36, 86–97.

Yuan, K., Srivastav, V., Navab, N., Padoy, N., 2024. Hecvl: Hierarchical video-language pretraining for zero-shot surgical phase recognition. arXiv preprint arXiv:2405.10075 .

Yuan, K., Srivastav, V., Tao, Y., Lavanchy, J.L., Mascagni, P., Navab, N., Padoy, N., 2023. Learning multi-modal representations by watching hundreds of surgical video lectures. arXiv.org doi:10.48550/arxiv.2307.15220.

Zhou, K., Yang, J., Loy, C.C., Liu, Z., 2022a. Conditional prompt learning for vision-language models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16816–16825.

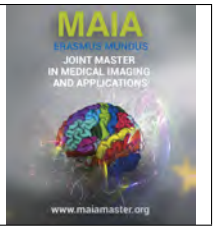
Zhou, K., Yang, J., Loy, C.C., Liu, Z., 2022b. Learning to prompt for vision-language models. International Journal of Computer Vision 130, 2337–2348.

Zhou, K., Yang, J., Loy, C.C., Liu, Z., 2022c. Learning to prompt for vision-language models. International Journal of Computer Vision 130, 2337–2348.

Ziogas, I.A., Tsoulfas, G., 2017. Advances and challenges in laparoscopic surgery in the management of hepatocellular carcinoma. World Journal of Gastrointestinal Surgery 9, 233.

Medical Imaging and Applications

Master Thesis, June 2024




# Multi-modal prediction of failed recanalization from pre-intervention neuroimaging (CT) and clinical data

Jesus Gonzalez, Pere Canals, Marc Ribo

*Stroke Research Unit, Vall d'Hebron Institut de Recerca, Barcelona, Spain*

**Abstract**

Effective prediction of failed recanalization in patients suffering from acute ischemic stroke (AIS) secondary to a large vessel occlusion could help improve outcomes in cases where conventional thrombectomy fails.

This study explores the potential of a multi-modal approach integrating various pre-intervention neuroimaging acquisitions (NCCT, CTA, CTP), clinical data and ground-truth segmentations as prior information to predict failed recanalization in non-cardioembolic stroke after first pass recanalization is not achieved, with an emphasis on cases involving intracranial atherosclerotic disease (ICAD). Using a dataset of 212 patients treated with endovascular therapy (EVT) at Vall d'Hebron Hospital, we implemented various machine learning models, including traditional algorithms and deep learning architectures. The models were trained on a combination of clinical variables, radiomic features, and imaging data, with additional experiments utilizing vessel and thrombus segmentation masks as prior information.

Our findings demonstrate that combining clinical and radiomic data significantly enhances predictive performance of tabular-based models compared to using either data source alone. On the other hand, the best multimodal model, a modified EfficientNet with a Dynamic Affine Feature Map Transform (DAFT) block to integrate imaging and clinical data, achieved an AUC of $0.74 \pm 0.12$, indicating robust predictive capabilities. Integrating vessel segmentation further improved model accuracy, underscoring the importance of multi-modal data fusion in predicting EVT outcomes.

Future research should focus on expanding the dataset and exploring additional data sources to further refine predictive models. The integration of advanced machine learning techniques and comprehensive data sources holds promise for enhancing stroke treatment protocols and patient outcomes.

*Keywords:* Acute Ischemic Stroke, Intracranial Atherosclerosis Disease, Failed Recanalization, Deep Learning, Multi-modal Classification

## 1. Introduction

A stroke is a severe neurovascular disease caused by various etiologies, broadly divided into two types: hemorrhagic and ischemic. Hemorrhagic stroke occurs when a blood vessel in the brain ruptures, leading to blood leakage, altered internal brain pressure, and severe neurological damage (Meschia, 2023). In contrast, ischemic stroke occurs when a blood vessel is occluded, disrupting the normal flow of blood to the brain cells. The blockage can be caused by a blood clot or thrombus, composed of various cell types such as cholesterol crystals, red blood cells or fibrin Jolugbo and Ariëns (2021). Ischemic strokes can be transient ischemic attacks (TIA) or acute ischemic strokes (AIS). While they may present similar symptoms, the former is a temporary event, and the latter can cause irreversible neurological damage caused by the death of brain cells in blood-deprived regions (Coutts, 2017). According to the most recent data from the Global Burden Disease Study (GBD), approximately 65% of strokes are ischemic, 28% are hemorrhagic, and the remaining are subarachnoid hemorrhages (IHME).

Among the main thrombus etiologies, we find cardioembolic (CE) occlusions, which occur when a thrombus originating in the heart due to a cardiac disorder embolizes, occluding a cerebral artery, or occlusions caused by intracranial atherosclerotic disease (ICAD). In patients suffering from ICAD, vessels are blocked by

plaque accumulated within the vessel walls, narrowing the vessel and eventually leading to flow disruption. Together, CE and ICAD occlusions account for approximately 45% of ischemic strokes. CE alone is responsible for 14-30% of ischemic strokes (Maida et al., 2020), while ICAD is highly prevalent in Asia and represents a significant portion of treatable strokes with EVT. In contrast, lacunar strokes, which occur in deep brain regions and account for 15-25% of ischemic strokes, are not typically treatable with EVT. In any case, a prompt response and treatment are essential, as time is critical in stroke cases (Saver, 2006).

### *1.1. Stroke Diagnosis*

When stroke symptoms are suspected in a patient, a rapid diagnosis-treatment protocol needs activation in the medical facilities where the patient is admitted. Treatment approaches depend heavily on the time from symptom onset to receiving medical attention.

Stroke diagnosis is first addressed in a pre-hospital phase undertaken by the emergency medical services. The in-hospital diagnosis pipeline primarily relies on brain imaging in accordance to the guidelines issued by the American Stroke Association and European Stroke Organization (Shulman and Abdalkader, 2023). Specific acquisitions may vary between medical centers according to policy and available resources.

Clinical assessment is made by the stroke neurologist and involves administering a set of motor, visual and verbal tasks tasks that, depending on the level of damage, can indicate the patient's stroke status. This is evaluated using the National Institutes of Health Stroke Scale (NIHSS), ranging from 0 to 42, with a higher value indicating a severe stroke (Meschia, 2023).

After clinical assessment, the imaging steps of the diagnosis pipeline follow. The decision between conducting a Computed Tomography (CT) or a Magnetic Resonance Imaging (MRI) depends on medical facilities and device availability. Non-Contrast Computed Tomography (NCCT) is mainly used to rule out or confirm hemorrhagic lesions, which appear hyper-intense in NCCT (Shulman and Abdalkader, 2023), and to assess the infarcted brain areas. The Alberta Stroke Program Early CT Score (ASPECTS) scale is employed to that end (Pexman et al., 2001).

Typically, Computed Tomography Angiography (CTA) is also acquired if a hemorrhagic stroke is ruled out in NCCT. In CTA, intra-venous contrast is administered to the patient allowing observation of the patient's blood vessels. CTA is commonly used for its high sensitivity in detecting large vessel occlusion (LVO) (Sanchez et al., 2024; Shulman and Abdalkader, 2023) and intracranial stenosis as stated in the Stroke Outcomes and Neuroimaging of Intracranial Atherosclerosis (SONIA) study, when compared to digital subtraction angiography DSA, making it a reliable non-invasive alternative, highlighting its role in both acute and chronic settings of stroke management(Sanchez et al., 2024).

In some centers, patients may undergo Computed Tomography Perfusion (CTP) to study brain blood hemodynamics (Wing and Markus, 2019). Depending on symptoms' severity and medical facility availability, patients may be imaged with an MRI, which is more precise but time-consuming, has more contraindications and is prone to motion artifacts.

### *1.2. Treatment for Ischemic Stroke*

Common AIS treatment methods include intravenous thrombolysis (IVT) and endovascular therapy (EVT). IVT involves injecting a thrombolytic agent, recombinant tissue plasminogen activator (rtPA), to the patient to dissolve the clot and restore blood flow. Its effectiveness is proven, but it is time-sensitive, depending on the time elapsed between symptom onset and rtPA administration (Grotta, 2023).

EVT or mechanical thrombectomy is an invasive procedure for AIS caused by a LVO. It involves inserting a catheter through a major artery to reach and extract the clot, recanalizing the affected blood vessel and establishing reperfusion. The success of this technique is significantly higher compared to using rtPA alone, as demonstrated by the HERMES trials in 2015 (Goyal et al., 2016), and both treatments can be used in combination if there are no contraindications (Phan et al., 2017). There is ongoing discussion on the treatment's effectiveness and whether there is a benefit to administering rtPA before EVT compared to practicing EVT alone (Phan et al., 2017).

EVT outcome depends on factors such as the thrombus's location within the brain's circulatory system, with more distal vessels being smaller and more fragile, making them prone to injury and rupture (Sheth, 2023). The success of EVT is also associated with thrombus composition, vessel geometry, and clot etiology, among other factors (Sheth, 2023).

After EVT, results can be assessed using the modified Thrombolysis in Cerebral Infarction (mTICI) scale evaluated on DSA series, indicating different levels of reperfusion. Values can range from 0 to 3, with intermediate levels 1, 2A, 2B, and 2C, with 3 representing complete reperfusion after the procedure and 0 indicating no reperfusion at all (Sheth, 2023). Patients with successful reperfusion are usually considered those with a reperfusion level greater than 2B. Reperfusion rates as high as 80-90% are reached in the latest trials (Fischer et al., 2022). Successful reperfusion is associated with better clinical outcomes and reduced mortality, making EVT the gold standard treatment for LVO (Beaman et al., 2022).

Other treatment options include contact aspiration, stent retriever thrombectomy, rescue angioplasty, and rescue stenting (Beaman et al., 2022):

- **Contact Aspiration:** This method involves using a suction device to aspirate and remove the thrombus directly.
- **Stent Retriever Thrombectomy:** A stent retriever is deployed at the site of the thrombus, trapping the clot and allowing it to be removed when the stent is retracted.
- **Rescue Angioplasty:** This technique uses a balloon to widen the blood vessel at the site of occlusion, which can help restore blood flow.
- **Rescue Stenting:** In cases where other methods fail, a stent can be permanently placed to keep the vessel open and maintain blood flow.

#### *1.2.1. Failed Recanalization*

Failed recanalization refers to the failure to recanalize the occluded vessel through IVT and/or EVT in patients suffering from LVO. EVT in patients with ICAD-LVO presents higher failed recanalization rates (Banerjee and Chimowitz, 2017; Rodrigo-Gisbert et al., 2024). This failure is associated with worse clinical outcomes, including higher rates of disability and death (Beaman et al., 2022).

#### *1.2.2. First Pass Effect*

In EVT, multiple attempts can be made to retrieve the thrombus causing the LVO. These are referred to as passes. Recanalizing after the first pass is commonly known as the First Pass Effect (FPE), which is associated with better clinical long-term outcomes and reduced progression of the ischemic lesion (Zaidat et al., 2018).

### *1.3. Imaging in Stroke diagnosis*

Once the assessment pipeline is defined and imaging is acquired, it is important to correctly interpret the information coming from both clinical assessments and imaging results.

#### *1.3.1. NCCT*

As mentioned earlier, NCCT is evaluated using the ASPECTS score (Pexman et al., 2001). This scale assesses 10 regions of interest in the brain at the ganglionic and supraganglionic levels. These regions include Caudate (C), lentiform nucleus (L), internal capsule (IC), insular ribbon (In), anterior Middle Circulatory (MCA) cortex (M1), MCA cortex lateral to the insula (M2), posterior MCA cortex (M3), anterior part of the MCA territory immediately superior to M1 (M4), lateral part of the MCA territory immediately superior to M2 (M5), and posterior part of the MCA territory immediately superior to M3 (M6) as seen in Figure 1 from Shulman and Abdalkader (2023). One point is assigned if no signs of early ischemic changes are detected and zero if they are. Once all regions are evaluated, their scores are summed, and the lower the result, the riskier the stroke.

An unequivocal sign of stroke in NCCT is the unilateral hyperdensity of a proximal large vessel, highly detectable within the lumen (Shulman and Abdalkader, 2023). When this sign is visible, that is not an specific sign, patients should be treated as if it were an LVO.

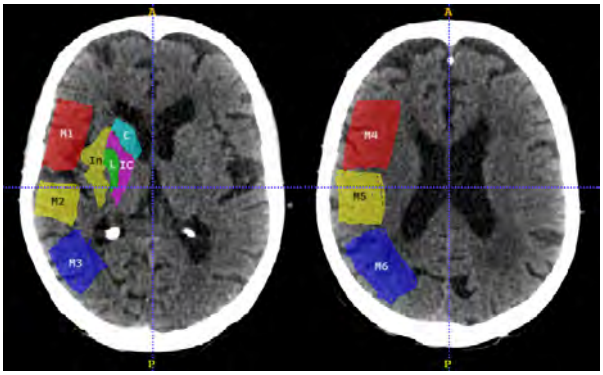



Figure 1: NCCT image example and ASPECTS region of interest.

#### *1.3.2. CTA*

CT angiography (CTA) is a highly accurate imaging technique used to detect large vessel occlusions (LVO) in the brain. CTA diagrams brain vessels using intra-venous contrast previously injected into the patient. The contrast in the lumen [1] of the extracranial and intracranial vasculature is captured and two-dimensional maximal-intensity projections and three-dimensional reconstructions can be reconstructed (Shulman and Abdalkader, 2023). In a normal blood flow scenario, the contrast dye used in CTA should spread evenly through the blood vessels, making them visible on the scan. An occlusion, or blockage, will appear as an area where the contrast dye does not reach, indicating a lack of contrast opacification. The importance of CTA lies in its extremely high accuracy, with a sensitivity of 98.4% and specificity of 98.1% in detecting LVOs (Shulman and Abdalkader, 2023).

CTA can be multiphase, acquiring images at different time points: peak arterial, middle, and late venous phases. This multiphase approach not only helps in assessing a large vessel occlusion (LVO) but also in evaluating collateral blood flow. Collateral flow refers to alternative or indirect arterial pathways that can potentially provide blood flow when an artery normally supplying an area of brain tissue is occluded. Understanding the status of collateral circulation is important because it can influence treatment decisions and outcomes in stroke patients. In quick diagnoses, it is common to

[1] The cavity or channel within a tube or tubular organ such as a blood vessel

use single-phase CTAs, which are faster but may not provide as comprehensive an assessment as multiphase CTAs (Dundamadappa et al., 2021; Yeo et al., 2001).

#### *1.3.3. Other Techniques*

CTP exposes the patient to more radiation than CTA, but it can be beneficial when measuring cerebral blood flow (CBF), cerebral blood volume (CBV), mean transit time, and approximations of the size and location of the infarct core and the surrounding damaged area, known as penumbra (Shulman and Abdalkader, 2023). Diffusion-weighted image-MRI has higher sensitivity and accuracy than NCCT but tends to be more complicated and time-consuming, making it less reliable in urgent cases. Another technique to check cerebrovascular issues is DSA. DSA is the gold standard for LVO location and hypoperfused region assessment, but it an invasive imaging method.

Magnetic resonance angiography (MRA) is a non-invasive technique that provides an accurate representation of the arterial lumen without radiation exposure. However, it generally has lower spatial resolution compared to DSA and CTA (Sanchez et al., 2024).

Transcranial ultrasound (US) is a non-invasive method that assesses blood flow velocity to infer stenosis. It provides hemodynamic information but does not directly visualize the plaque and is highly operator-dependent (Sanchez et al., 2024).

Optical coherence tomography (OCT) uses an intravascular probe to provide high-resolution images of plaque characteristics such as intimal thickening and lipid accumulation. Despite being invasive, OCT offers detailed assessment of plaque morphology, useful for understanding plaque stability (Sanchez et al., 2024).

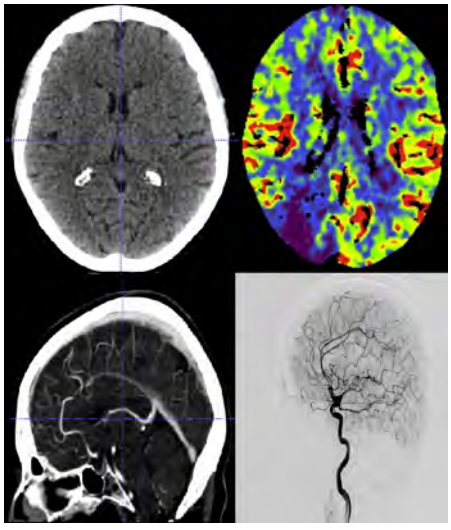


Figure 2: Different types of images: from left to right and top to bottom: NCCT, CTP, CTA, DSA.

### *1.4. ICAD*

Atherosclerotic plaque can be defined as a buildup of fibrin and lipid tissue within the intracranial arterial walls (Beaman et al., 2022) and is one of the most common causes of stroke worldwide (Banerjee and Chimowitz, 2017), highly prevalent in Asia, representing around 50% of all stroke cases and 10% in the United States of America. Since it affects Hispanics and Afro-Americans more sensibly than Caucasians, major drivers of population growth, it is expected that ICAD-LVO incidence cases will rise over time (Banerjee and Chimowitz, 2017).

Treatment of ICAD-LVO includes IVT and EVT, along with the same diagnosis pipeline. Nevertheless, despite EVT being the gold standard for patients with underlying ICAD, this disease has been associated with lower recanalization rates (Rodrigo-Gisbert et al., 2023), longer procedural times, cognitive decline, and increased global economic burden (Beaman et al., 2022) compared to CE-LVO.

The imaging modality commonly used to diagnose ICAD is CTA, which is the most accurate non-invasive method with high specificity and sensitivity(Banerjee and Chimowitz, 2017). While CTA is precise in identifying occlusions and can also detect stenoses, it generally cannot determine whether an occlusion is due to intracranial atherosclerotic disease (ICAD). The gold standard for identifying ICAD is DSA, which provides more detailed information, though even DSA can have variability in identifying plaque composition and stenoses. Clinical history, such as the NIHSS score, which tends to be lower than expected, can provide additional context (Beaman et al., 2022).

Diagnosing ICAD-LVO is particularly challenging mostly due to the small amount of cases and also due to the presence of imitators such as intracranial vasospasm, dissection, and partially occlusive thrombus, which can mimic the appearance of ICAD-LVO during imaging and mechanical thrombectomy (Rodriguez-Calienes et al., 2024). These conditions can lead to misdiagnosis and inappropriate interventions. For example, intracranial vasospasm can appear similar to ICAD-LVO on imaging but is typically reversible, whereas ICAD is persistent. Distinguishing between these conditions is critical to avoid unnecessary and potentially harmful treatments.

When recanalization fails during EVT, the decision to continue with additional passes or to use a rescue treatment, such as stenting or angioplasty, depends largely on the discretion and expertise of the interventionist in situ. This highlights the complexity and difficulty of managing ICAD-LVO, even for experienced clinicians.

### *1.5. Our Work*

It is our interest to timely identify potential failed recanalization from pre-operational images, particularly

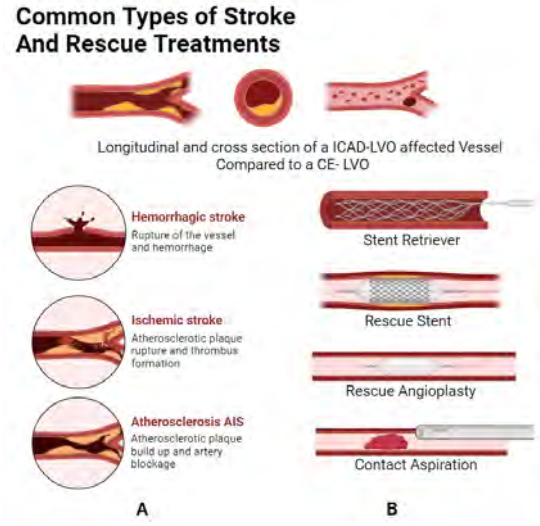



Figure 3: A: Intracerebral Hemorrhage (ICH), Cardio Embolic LVO and Intracranial atherosclerosis disease related LVO. B: EVT types and devices

in ICAD-related LVO, to provide useful insights for appropriate treatment. For example, early confirmation of ICAD-LVO may reduce the number of stent retriever passes, thereby reducing vessel damage and facilitating proper ICAD procedures such as angioplasty and stenting (Haussen et al., 2018).

As part of our work, we aim to predict ICAD using a combination of tabular data and imaging data. The tabular data consists of clinical variables, imaging-derived variables, and radiomic features from thrombus images. We will explore the predictive capabilities of each type of data individually—clinical variables alone, radiomic features alone—and in combination.

To set a baseline, we will conduct experiments using imaging data from NCCT and CTA scans, both separately and together. These experiments will be performed on the whole 3D volume as well as on skull-stripped images. Skull stripping reduces the information in the image by keeping only the brain, whereas using the complete volume includes parts of the vascular system that may show stenosis or ICAD, which could be informative for our prediction.

Based on the best results between skull-stripped and complete volumes, we will integrate tabular data and imaging data to examine if the inclusion of more data improves the prediction outcomes.

Developing an accurate discriminator is desired, as it could enable clinical trials to test different stroke rescue treatments, such as stenting, beyond just increasing the number of mechanical thrombectomy passes. This approach is important because stenting can be beneficial for ICAD but may be more harmful than helpful for CE. Therefore, accurately distinguishing between these conditions is necessary to avoid potential harm or unnecessary costs associated with stenting everyone indiscriminately.

To the best of our knowledge, while there has been some work using deep learning in MRA, there is no record of a study that combines traditional machine learning methods, deep learning, and multimodal integration for ICAD-LVO using pre-operative CT images. Our work aims to bridge this gap by exploring the integration of various data modalities to improve detection and classification outcomes in ICAD-LVO.

The rest of this document is organized as follows: a review of the state of the art in Chapter 2, explaining the nature of the problem; Chapter 3 details the data used and the experiments conducted on different data modalities to set a baseline, as well as the multimodal approach. The results section shows metrics for both baseline and final experiments, followed by a discussion and, finally, a conclusion.

## 2. State of the art

### *2.1. Clinical Predictors*

Efforts to diagnose LVO with underlying ICAD early are motivated by the potential to improve treatment strategies. Early identification of recanalization likelihood allows for anticipating procedural strategies, saving time, and reducing complications. Although EVT is highly effective, it is particularly challenging in ICAD-LVO cases due to high risks of re-occlusion and permanent damage (Cai et al., 2022; Haussen et al., 2018). Thus, swift diagnosis is important for selecting the appropriate EVT approach (Li et al., 2022).

#### *2.1.1. Clinical Variables*

Several clinical variables are linked to ICAD-LVO, different studies have confirmed the different associations of different predictors with ICAD-LVO Haussen et al. (2018); Li et al. (2022); Liao et al. (2022); Rodrigo-Gisbert et al. (2024); Zha et al. (2021). Established clinical predictors include:

- High HbA1c levels
- Presence of LDL cholesterol or high dyslipidemia
- Elevated Systolic Blood Pressure
- Absence of atrial fibrillation
- Hypertension

#### *2.1.2. Hyperacute Clinical Biomarkers*

Markers assessed during the acute phase provide additional insights:

- Baseline NIHSS: Patients with ICAD-LVO often have milder NIHSS scores compared to other LVO types (Psychogios et al., 2022).

- ASPECTS: Higher scores in baseline NCCTs are typically associated with ICAD-LVO (Chen et al., 2023).
- Onset-to-image (OTI) time: Used alongside CTP parameters to assess collateral status.

#### *2.1.3. CTP Image-Derived Markers*

Markers derived from CTP imaging relevant to ICAD include:

- Core infarct volume (CBF < 30%): Reflects the blood flow through brain tissue. Values below 30% of normal flow, compared to healthy brain areas, indicate irreversibly infarcted volume (Amukotuwa et al., 2019).
- Tmax>4/Tmax>6 ratio: Tmax represents the time it takes for blood to arrive at a given region of the brain. A Tmax>4 indicates areas with significant delays, and a ratio above 2 suggests good collateral flow (Haussen et al., 2018).
- Hypoperfusion Intensity Ratio (HIR): Defined as Tmax>10/Tmax>6, where a high ratio indicates poor collateral status. HIR helps in identifying the severity of perfusion deficits (Lyndon et al., 2021).
- Tmax>4: Known as Hypoperfused volume growth rate, it measures the volume of tissue experiencing delayed blood flow and serves as a predictor for ICAD (Rodrigo-Gisbert et al., 2024).

An important CTP-related predictor is the Hypoperfusion Intensity Ratio (HIR), defined as Tmax>10s/Tmax>6s. HIR has gained relevance in stroke prognosis as studies have confirmed that a low HIR, in combination with other known predictors, suggests underlying ICAD (Rodrigo-Gisbert et al., 2024). Patients with good collateral circulation are considered slow progressors since collateral flow sustains the brain tissue, slowing the ischemic core's advancement. A high HIR indicates a large ischemic penumbra, which is the area that can potentially be saved by recanalization. ICAD patients tend to have good collaterals due to the progressive nature of the disease, which causes small intracranial stenoses and longstanding ischemic regions, leading to enhanced collateral flow (Maguida and Shuaib, 2023).

#### *2.1.4. Image-Derived Markers*

Additional measurable markers from imaging studies include:

- Hyperdense sign (HS) in NCCT: Indicates CE thrombus but is absent in about 30% of all clots.
- Truncal occlusion type: More commonly associated with ICAD-LVO, compared to branch occlusions typical of CE occlusions.
- Delta HU: Small HU differences between the thrombus region and a contralateral patch can indicate ICAD-LVO (Siddiqui et al., 2023).
- HU ratio: A ratio close to 1 between thrombus patch intensity and the contralateral patch is indicative of ICAD-LVO (Siddiqui et al., 2023).

#### *2.1.5. Other Image-Derived Markers*

General imaging markers aiding in predicting ICAD-LVO include:

- Calcifications: Presence of intracranial calcifications suggests ICAD.
- Collaterals: Good collateral circulation is a significant indicator of ICAD-LVO (Maguida and Shuaib, 2023).
- Atherosclerosis in other regions: Atheromatosis in areas such as the carotid bifurcation or aortic arch can indicate ICAD.
- Thrombus Radiomics: Thrombus radiomics involves extracting detailed features from CT images of thrombi to predict outcomes and complications during EVT van Voorst et al. (2023). For example, Yusuying et al. (2023) developed a CT-based thrombus radiomics nomogram to predict secondary embolization (SE) during EVT for LVO. The study extracted 107 radiomics features from pre-interventional CT images, including first-order statistics, shape-based features, and texture features such as gray-level co-occurrence matrix (GLCM) and gray-level run length matrix (GLRLM). These features were used to develop a support vector machine (SVM) learning model that demonstrated high predictive accuracy.

### *2.2. Evaluation Scales*

Custom evaluation scales have been proposed to predict ICAD-LVO using pre-intervention data:

- Zha et al. (2021) proposed a scale based on patient history of hypertension, atrial fibrillation rhythm, and baseline serum glucose, with scores ranging from -4 to 4. This predictive scale, referred to as the ISAT scale, was developed to predict in situ atherosclerotic thrombosis in acute vertebrobasilar artery occlusion (VBAO) patients before EVT. The scale was validated using a derivation cohort from the Nanjing Stroke Registry Program and an external validation cohort, showing good discrimination with an area under the receiver operating characteristic curve (AUC) of 0.853 in the derivation cohort and 0.800 in the validation cohort.

- Liao et al. (2022) developed the ABC2D score to predict the etiology of intracranial LVO before EVT. The score incorporates atrial fibrillation, blood pressure, clinical neurological deficit, the CT hyperdense sign, and diabetes mellitus. The ABC2D score was derived and validated in a large cohort, demonstrating high predictive value with AUC values of 0.886 and 0.880 in the derivation and validation cohorts, respectively.

- Chen et al. (2023) introduced the ATHE scale, which includes the absence of atrial fibrillation, the presence of truncal-type occlusion, the absence of a hyperdense artery sign, and a lower baseline examination NIHSS score as key predictors of ICAD-LVO. This scale was developed to identify the most significant predictors and then validated through logistic regression. The ATHE scale demonstrated excellent predictive performance with an AUC of 0.920 in the derivation cohort and 0.890 in the external validation cohort.

### *2.3. Prediction Models using tabular data*

We have reviewed various predictive variables for identifying ICAD-LVO before EVT. These variables include clinical data, hyperacute indicators, CTP-derived metrics, and imaging-derived features. In this section, we will present different models that use these variables, either individually or in combination, to predict ICAD-LVO. These models use tabular data to systematically analyze and interpret the complex interactions between the predictive variables, ultimately aiming to improve decision-making processes in clinical settings.

Several models are utilized when using one predictor or a combination of predictors to identify ICAD-LVO. Typically, these models employ univariate logistic regression (LR) or multivariate logistic regression (MLR), as supported by studies such as Cai et al. (2022); Haussen et al. (2018); Liao et al. (2022); Rodrigo-Gisbert et al. (2023).

However, advanced techniques like Random Forest are also employed, as demonstrated by the study of van Voorst et al. (2023), which includes radiomics features specifically for ICAD-LVO pre EVT detection. Other studies, like Yusuying et al. (2023) have experimented with various models such as LR, support vector machine (SVM), K nearest neighbor (KNN), random forest (RF), extremely randomized trees (Extra-Trees), eXtreme Gradient Boosting (XGBoost), light gradient boosting machine (LightGBM), and multilayer perceptron (MLP), ultimately identifying the SVM model as having the highest average area under the receiver operating characteristic (ROC) curve (AUC) for predicting the risk of secondary embolization. Although this study is notable, it focuses on LVO without considering ICAD.

Use of SVM is also seen in the work done by Bento et al. (2019), where they use an SVM classifier over a wide image features set extracted from magnetic resonance imaging sequences, setting a multiclass classification problem among 4 different classes: carotid artery atherosclerostic disease, multiple sclerosis, small vessel disease and normal controls.

### *2.4. Deep Learning Approaches*

Besides exploring classification tasks of several diseases and use of SVM, Bento et al. (2019) use a wide set of features including image intensity gradient-based attributes, local binary patterns, and frequency domain features. Deep learning models, particularly convolutional neural networks (CNNs), inherently extract high-level features from raw imaging data upon training, capturing complex patterns and structures that may be indicative of specific conditions. These features can then be leveraged for various tasks such as classification, segmentation or anomaly detection (Bento et al., 2018).

In stroke imaging, deep learning approaches have repeatedly been proposed for LVO detection, regardless of the occlusion's nature (Cui et al., 2022). These methods often employ either a single imaging modality or multimodal approaches, integrating NCCT, CTA, and CTP, and sometimes including clinical data in the form of previously described biomarkers.

Recent work has also focused on using deep learning for MRA to detect intracranial arterial stenosis and occlusion. For example, a study utilized the YOLOv5 detection model on time-of-flight MRA (TOF-MRA) images, achieving promising results in the automated detection of steno-occlusive lesions. This approach showed a sensitivity of 64.2% and a positive predictive value of 83.7%, particularly excelling in detecting lesions in the internal carotid artery (Qiu et al., 2022). Despite these advancements, there remain significant challenges in achieving high accuracy and consistency across different vascular territories and stenosis categories.

Additionally, there have been various challenges in LVO classification and ischemic lesion segmentation. Notable examples include the ISLES ischemic lesion segmentation challenge and the Image Analysis for CTA Endovascular Stroke Therapy (IACTA-EST) Data Challenge [2], which focused on LVO/no LVO classification. Interestingly, while there were good results using CTA for the IACTA-EST challenge, no significant success was achieved with multimodal tabular-imaging data, reflecting the challenging nature of such tasks. MICCAI 2024 will mark a significant milestone by hosting the first known intracranial stenosis challenge (INSTED)[3], pioneering efforts in this critical area and

[2] https://lgiancauth.github.io/iacta-est-2023/data-info

[3] https://miccai.org/index.php/special-interest-groups/challenges/miccai-registered-challenges/

underscoring the evolving focus on intracranial stenosis detection and classification.

## 3. Material and methods

### *3.1. Dataset*

The dataset used for this study was collected from cases at Vall d'Hebron Hospital between January 2018 and December 2022. Initially, cases were considered based on the inclusion criteria of the site of the LVO: TICA, M1, and M2, and having received EVT. Patients with bilateral and chronic occlusions were excluded.

However, not all cases met the necessary criteria for inclusion, rendering them non-viable for the study. An image revision process was conducted to exclude images with occlusions outside the considered brain locations, images with incorrigible artifacts, no occlusions, or bilateral occlusions.

An in-depth review of patient history was performed to remove patients without ICAD/recanalization information, those with occlusions in excluded locations, or those whose initial eTICI value indicated good reperfusion. The data selection process is illustrated in Figure 4.

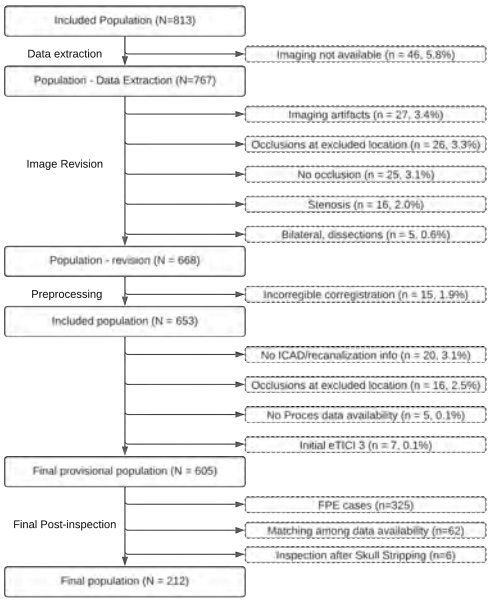



Figure 4: Data Selection Process

Given that the primary objective was to identify patients with failed recanalization using conventional treatment, all cases regardless of the failure reason were initially included. However, several considerations were noted:

- A large fraction of cases with failed recanalization had an occlusion from a cardioembolic (CE) source. AIS due to CE thrombi generally have better recanalization rates than other etiologies like ICAD (Bang et al., 2010), and we hypothesize that these can contribute to heterogeneity in the group of failed recanalizations.
- Stroke etiology was indeterminate in some cases.
- A small portion of the cases with recanalization success were due to rescue stenting.

To improve the homogeneity and relevance of the dataset, target group was limited to the following population:

1. **ICAD**: cases with confirmed intracranial atherosclerotic disease (ICAD) were particularly interesting due to their lower recanalization rates and higher recurrence tendency. Hence, all ICAD cases were included, even if they achieved successful reperfusion, to explore the potential benefits of intracranial stenting as part of the conventional treatment for these cases[4].
2. **Failed Recanalization with Indeterminate Etiology**: cases with indeterminate etiology were included because identifying the underlying cause of recanalization failure in these cases could provide valuable insights.
3. **Rescue Treatment Cases**: all cases that required rescue treatment were considered of interest as they indicate failure of conventional EVT and present an opportunity to explore alternative interventions.

In summary, the final positive class in the dataset includes all cases with failed recanalization using EVT, excluding CE cases. Additionally, it includes all ICAD cases regardless of recanalization success due to their high recurrence rates and the potential need for stenting even after initial recanalization.

The goal of the current classification problem is to identify cases that will not recanalize or are ICAD. However, based on the current available evidence, regardless of the model's prediction, it is reasonable to establish that all cases will undergo at least one EVT pass in a realistic setting. A significant number achieve First Pass Effect (FPE). Therefore, cases that achieve FPE were excluded in order to select a representative population in our development.

As the study's objective is to explore multimodal prediction, a final data reduction was performed by including cases with complete imaging data (NCCT and CTA), CTP information, a complete set of clinical variables (demographics, risk factors, and hyperacute variables), and radiomics data from previous manual thrombus segmentation. Additionally, vascular segmentation extracted using Arterial (Canals et al., 2023) and thrombus manual annotation segmentation were included. The resulting dataset from the overlap of the different data availability is illustrated in the Venn diagram in Figure 5.

[4] After at least one thrombectomy pass

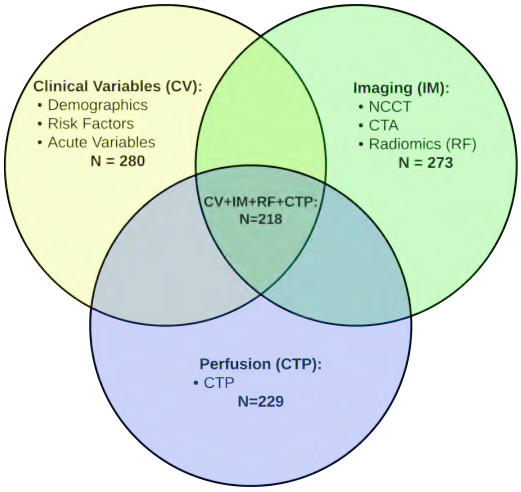



Figure 5: Venn diagram showing the overlap of cases with complete imaging, clinical, and radiomics data for the final dataset.

The final used dataset comprised cases that met all the inclusion criteria, were not excluded based on the above considerations, and had complete data availability for the study's multimodal prediction analysis.

### *3.1.1. Imaging Data Preprocessing*

The initial image format was Digital Imaging and Communications in Medicine (DICOM). The dataset was converted to the more compact Neuroimaging Informatics Technology Initiative (NiFTI) format.

Images were registered to a common space for simplicity. First, NCCT images were registered to CTA, then CTA images were registered to MNI space, and that transformation was applied to the NCCT images. As seen in Figure 4, the registration process was reviewed to eliminate samples with incorrigible registration errors.

Initially, the images had a resolution of 421x505x452 with a voxel size of 0.43x0.43x0.4 mm. During the initial stages of model training, it was observed that the original image size caused errors due to computational resource limitations. After consulting with the medical and engineering teams, it was suggested to change the voxel size to a value smaller than 1mm but larger than the original size, in order to preserve lesions that are typically smaller than 1mm. The new voxel size was set to 0.73x0.73x0.7 mm.

Initially, the images were resized to approximately half the size in each dimension to a resolution of 210x250x226 and then voxel-resized, resulting in final dimensions of 153x182x158. However, the results at this size were not satisfactory. Considering that voxel spacing implicitly reduces the image dimensions, resizing was applied directly to the original images, resulting in final dimensions of 246x295x258. This resizing step was necessary to ensure the feasibility of the training process given the available computational resources and was applied to NCCT, CTA vascular segmentation, and thrombus segmentation files.

Skull stripping process was carried out using TotalSegmentator (Wasserthal et al., 2023) and was applied only to CTA and NCCT images. After a visual inspection of the skull stripping process, some samples were removed due to image orientation errors, potentially derived from incorrigible registration errors.

Table 1: Dataset Dimensions and Voxel Sizes

| Dataset | Resolution | Voxel Size (mm) |
|---|---|---|
| Original Dataset | 421x505x452 | 0.43x0.43x0.4 |
| Resampled | 246x295x258 | 0.73x0.73x0.7 |
| Resampled-stripped | 246x295x258 | 0.73x0.73x0.7 |

Further inspection of the images reveals that NCCT ranges from -1260.84 ± 60.90 to 2093.80 ± 252.54 Hounsfield Units (HU) and CTA from -1292.32 ± 41.07 to 2477.80 ± 246.57 HU. These ranges were clipped to (0, 100) HU for NCCT and (0, 400) HU for CTA in order to strengthen the visibility of regions of interest. This clipping is a common preprocessing technique (Patel, 2023). Furthermore, images were scaled from 0 to 1 as it is a good practice for subsequent deep learning treatment (Montavon et al., 2012).

## *3.2. Methodology*

This study is organized into three main stages. First, different combinations of available predictors, including clinical variables, radiomic features and CTP parameters, are tested using different classical machine learning algorithms to set baseline scores and study the predictability of the different considered subsets. Second, an imaging baseline is built using the preprocessed CT volumes through different deep convolutional neural networks. Models were trained using NCCT, CTA or both. Third, imaging models are used alongside clinical predictors in order to study a potential enhancement of the results.

### *3.2.1. Baseline Experiments with Tabular Data*

Data subsets in this section are as follows: one dataset (clinical) comprises clinical variables, hyperacute variables, and CTP-derived biomarkers: age, sex, hypertension (HT), dyslipidemia (DL), diabetes mellitus (DM), atrial fibrillation (AF), smoking (SM), NIHSS Baseline, wake-up stroke, site of the occlusion: side, TICA, proximal/distal M1 or proximal/distal M2, and IVT. From CTP: CBF<30%, Tmax>10s, Tmax>6s, Tmax>4s, HIR, and Tmax4s/Tmax6s.

The second subset (radiomics) includes a comprehensive set of radiomic thrombus features derived not only from the manual segmentation on NCCT but also from CTA images. Additionally, features were extracted not

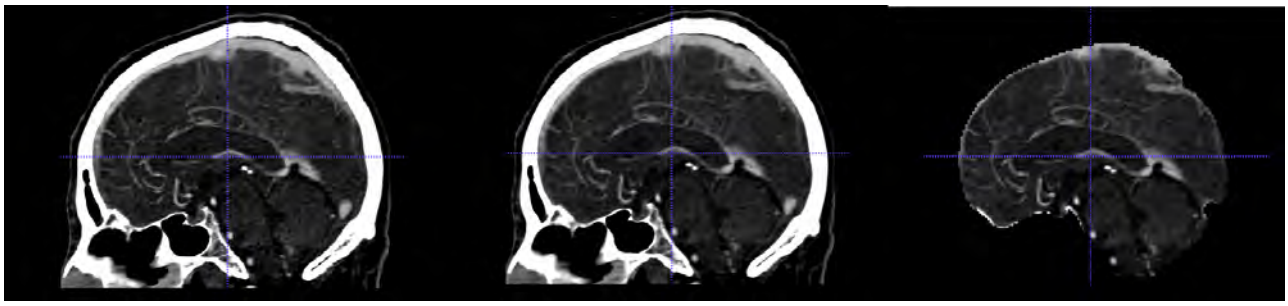


Figure 6: Dataset Dimensions - CTA Sagittal View. From left to right: Original Dataset, Resampled, Resampled-stripped

only from the occlusion site but also from the contralateral patch of the occlusion site (Lal-trehan and Giancardo, 2021; Siddiqui et al., 2023). The total number of predictors in this dataset is 5,168.

The final subset (clinical+radiomics) is the combination of the previous two datasets.

In the baseline experiments, various classical machine learning algorithms were employed to evaluate the predictive performance of the different datasets. A set of those algorithms is included in the Scikit-Learn Python library (Pedregosa et al., 2011):

- Support Vector Machine (SVM) is a supervised learning model that classifies data by finding the optimal hyperplane that separates data points of different classes.
- Random Forest (RF), an ensemble learning method, constructs multiple decision trees during training and outputs the mode of the classes for classification.
- Gradient Boosting (GB) is another ensemble technique that builds models sequentially, with each model attempting to correct the errors of the previous one.
- Logistic Regression (LR) is a statistical model that predicts the probability of a binary outcome based on one or more predictor variables.
- Lastly, the Multilayer Perceptron (MLP) is a class of feedforward artificial neural networks that consists of at least three layers of nodes.

The rest are included in the xgboost python library (Chen and Guestrin, 2016):

- XGBoost (XGB) is an optimized gradient boosting framework that uses tree-based learning algorithms.
- XGBoost with Random Forest (XGBRF) Classifier combines the strengths of XGBoost and Random Forest for improved performance.

The experimental setup involved several steps. First, an exploratory data analysis (EDA) phase was conducted for the clinical subset, involving typical data cleaning, verification, and checking for missing values to ensure data quality. Each dataset was then subdivided into three versions: the original dataset without any transformation, the dataset scaled to the 0-1 range for normalization, and the dataset scaled and passed through a Recursive Feature Elimination (RFE) procedure using Random Forest. RFE is a feature selection method that recursively removes the least important features based on model performance (Guyon et al., 2002).

The procedure for radiomics data followed the same approach as for clinical data but included an additional Recursive Feature Elimination (RFE) step due to the large number of available predictors. For the combined clinical and radiomics data, the RFE procedure was performed separately on the radiomics data before concatenating it with the clinical data.

Models were evaluated in a default parameter fashion and the top-performing model on default parameters underwent extensive hyperparameter tuning to further improve results. In both cases, the weights parameter was set so each model (where possible) could consider the weights of each class, which means considering the proportion of each class. This parameter was found useful considering how imbalanced the problem is.

Metrics used were AUC, F1 score, weighted accuracy, and confusion matrix. These metrics help assess model performance, especially in terms of balancing false positives and false negatives. 5CV approach was employed to ensure robust performance evaluation in both default parameter settings and during hyperparameter tuning.

After training the models, feature importance was assessed using SHAP library (SHapley Additive exPlanations) (Lundberg and Lee, 2017) force plots and summary plots to understand how individual features contribute to the model's prediction for each data instance, benefiting from the fact that SHAP is model agnostic.

#### *3.2.2. Experiments with Image Data*

Image data preprocessing was performed as explained in section 3.1.1. Training was conducted

using three different baseline models: ResNet34, DenseNet169, and EfficientNetB0 (EffB0). DenseNet is an evolution of ResNet, designed to improve information and gradient flow through dense connections between layers. EfficientNet further builds on DenseNet by optimizing both the depth and width of the network for better performance and efficiency (Tan and Le, 2020). Each model incorporates different architectural components that may be beneficial for the task at hand; for instance, ResNet uses residual connections to mitigate the vanishing gradient problem (He et al., 2015), DenseNet uses dense connections to enhance feature reuse (Huang et al., 2018), and EfficientNet uses a compound scaling method to uniformly scale all dimensions of depth, width, and resolution.

Each model was trained with both resampled images and resampled and skull-stripped images. Training was conducted using just NCCT volumes, just CTA volumes, and by stacking both images as different channels of the same image. Additionally, segmentation mask information, specifically vessels and thrombi, was used as prior information. These masks were used as additional channels, either the vessel, the thrombus, or both, in the same training fashion: prior information on NCCT, on CTA, and on both.

Figure 7: Example of prior information segmentation mask for a case with a left distal M1 occlusion. Vessels (red) and thrombus (green) masks.

To address the problem of small datasets, data augmentation techniques were applied. Data augmentation is a known method that helps to balance datasets by artificially increasing the size and variability of the training data. Various augmentation methods were tested, but the most effective ones retained were 'RandFlip' and 'RandZoom'. The 'RandFlip' method randomly flips the images along the specified spatial axis, providing mirrored versions of the original images, which helps the model generalize better by learning from different orientations. The 'RandZoom' method randomly zooms in and out on the images within a specified range (0.9 to 1.1), introducing slight variations in scale that help the model become more robust to differences in image size.

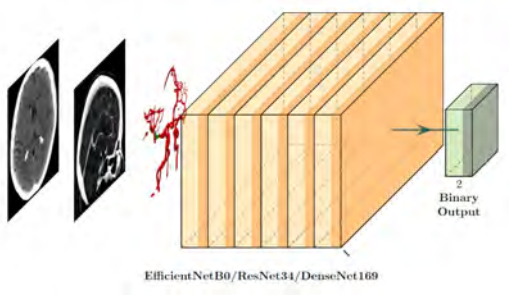



Figure 8: General Scheme of the Imaging Experiments. Input images (and the combinations of them) go through each of the different backbone architectures

All networks were trained from scratch, meaning no pre-trained networks were used. The models were implemented using the MONAI medical imaging framework (Cardoso et al., 2022) and all training procedures employed 4-fold cross-validation (4CV).

### *3.2.3. Multimodal Integration Experiments*

In the context of this experiment, an intermediate data fusion approach was used. The logic behind the method involves using EffB0 as a feature extractor. EfficientNetBNFeatures, part of the MONAI medical imaging framework (Cardoso et al., 2022), was chosen due to its optimized architecture that scales depth, width, and resolution uniformly for better performance and efficiency.

The clinical information was integrated using a Dynamic Affine Feature Map Transform (DAFT) block-Polsterl et al. (2021). DAFT dynamically rescales and shifts the feature maps of a convolutional layer based on the patient's clinical information, effectively integrating high-dimensional image data with low-dimensional tabular data. This block is a general-purpose module that enhances the interaction between image and tabular data within the network.

To implement this, the EfficientNetB0Features model was adapted to serve as a feature extractor. The clinical information was then added using the DAFT block. This step required modifications to ensure the proper size and output compatibility between the EfficientNetB0Features and the DAFT block input, considering that the initial DAFT method uses ResNet as the feature extractor (Polsterl et al., 2021). Specifically, the DAFT block includes global average pooling of the image feature map, concatenation with tabular data, and a series of fully connected layers to generate scaling and shifting parameters for the feature maps.

The final model architecture (DaftEffB0) as seen in

figure 8 included modified additional layers that received as input the combined output from the DAFT block. These layers consisted of 3D convolutions and pooling layers, ending with a linear classifier to generate the final prediction. Batch normalization was added to the output of every convolutional layer to enhance model performance and stabilize training.

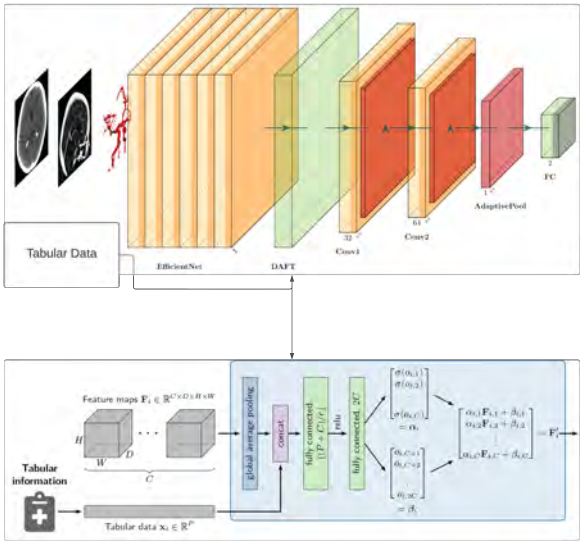



Figure 9: DaftEffB0 (top) with description of DAFT block by Polsterl et al. (2021) (bottom)

The training phase involved several configurations. Each model was trained as in the previous phase: using NCCT volumes, CTA volumes, and a combination of both. Additionally, the available segmentation masks: either individually or combined, for both NCCT and CTA volumes. This part training procedures employed again 4CV.

The metrics used for evaluating the multimodal integration experiments were consistent with those used for the image-only experiments, including Area Under the Curve (AUC), F1 score, weighted accuracy, and confusion matrix.

Additional parameters for training the deep convolutional networks could be set by the user, including the number of epochs, the name of the class of interest, the learning rate, and the use of the ReduceLROnPlateau scheduler. The ReduceLROnPlateau scheduler adjusts the learning rate dynamically based on the validation performance, helping to prevent overfitting and ensuring that the model converges more effectively by reducing the learning rate when a plateau in performance is detected. Other parameters included the model name, the imaging modality (NCCT, CTA, or both), the inclusion of prior information (vessel, thrombus, or both), input and output directories, and whether tabular data was used, including the path to the tabular data and the batch size.

For the imaging and multimodal experiments, the parameters were set as follows: batch size was 1, and the learning rate was 0.0001, adjusted using the ReduceLROnPlateau learning rate scheduler. The Adam optimizer was employed, with the loss function being Cross Entropy. The number of epochs varied between 40 and 60 per fold, contingent on the GPU capabilities, with higher channel images requiring longer training times.

In terms of hardware, a cloud service was utilized, offering a range of Nvidia GPU options as seen in table 2. The a6000 GPU allowed for a larger batch size due to its superior computational power. However, as a cloud service, the experiments were subject to the availability of the specific GPU models. The training phase was conducted using PyTorch, a widely-used deep learning framework. The custom software developed for this study saved fold indices, metrics, and model weights, facilitating the continuation of training, reproducibility of results, and further analysis if needed.

Table 2: Hardware Specifications

| Card | RAM (GB) | #CPU | GPU (GB) |
|---|---|---|---|
| P5000 | 30 | 8 | 16 |
| A4000 | 45 | 8 | 16 |
| RTX5000 | 30 | 8 | 16 |
| A6000 | 45 | 8 | 48 |

### *3.3. Model Ensemble*

To enhance the robustness and accuracy of our predictive models, we employed an ensemble model approach. This involved integrating predictions from multiple models trained on different imaging inputs (NCCT, CTA, and both) and using various prior information methods (vessel, thrombus, and both). Each of these models utilized the DAFT architecture and sampling methods for class balancing.

The ensemble model was constructed using a weighted majority voting scheme. This method combines the predictions from the individual models, weighting them according to their performance on true positive and true negative rates. The weights were calculated based on the balanced rates, adjusting for the class imbalance inherent in our dataset. This approach ensures that the most reliable models have a greater influence on the final prediction.

The weighted majority voting process involved aggregating the predictions from each model and applying the calculated weights to determine the final prediction. This method improves overall predictive performance by leveraging the strengths of each individual model (Dietterich, 2000). The combined approach mitigates the risk of overfitting and enhances the generalization capabilities of the model, making it more robust against class imbalance.

All tested models based on ensembling are Daft-EffB0, resampled and use sampling as the class balancing method.

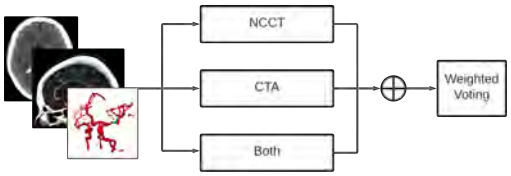



Figure 10: Ensemble weighted majority voting diagram

## 4. Results

Out of the 813 screened patients, 605 fulfilled the initial inclusion criteria and undergo necessary medical history revision for analysis. FPE (non-ICAD) was achieved in 325 (53.7%), and 68 were finally disregarded due to complete data unavailability (62, 10.2%) and incorregible preprocessing errors (6, 1.0%). The study finally included a total of 212 patients. The mean age of the patients was 77 ± 15 years, and 60.9% were women. The baseline NIHSS score was 15 ± 6. Among the patients, 36.32% received intravenous thrombolysis. Regarding the affected side, 52.8% had the LVO on the left side.

Focusing on the class of interest, there were 27 patients (12.7%). In this subgroup, the mean age was 74 ± 15 years. The baseline NIHSS score was lower (14 ± 6) compared to the complete sample. Within this class, 66.7% were women and 63.0% had a left-sided stroke. Additionally, 33.3% of the patients in this class received IVT. Among the class of interest, 9 patients (33.3%) were diagnosed with ICAD, and 5 patients (18.5%) received a stent as rescue treatment. In the total cohort, this represents 4.3% and 2.4% of the sample, respectively. The remaining patients (13, 48.1%) were cases where recanalization was not achieved and etiology of the stroke was undetermined.

The idea of the population belonging to the class of interest are the ones who could potentially benefit the most from alternative treatments to conventional methods. The detailed results of the experiments will now be presented in the following subsections.

### *4.1. Tabular Data*

Different tabular data experiments were conducted following the data description outlined in section 3.2.1.

The results of the experiments with clinical data and perfusion parameters are shown in Table 3. Each dataset was evaluated in two versions: data-as-is and scaled data. The results are presented as mean ± standard deviation, considering the different iterations in 5CV.

The experiments revealed that the integration of radiomics data significantly improved the model's performance compared to using clinical data alone. Specifically, the radiomics dataset achieved better AUC, F1-score and weighted accuracy. This improvement highlights the importance of structural and textural information captured by radiomics in understanding thrombus pathology and treatment response.

Further enhancement was observed when combining clinical and radiomics data. The combined dataset achieved the highest performance metrics. These results demonstrate that integrating multiple data sources can significantly enhance predictive performance. The high sensitivity indicates the model's effectiveness in correctly identifying positive cases, although the specificity was somewhat lower. This trade-off suggests that while the model is highly effective at detecting true positives, it may also produce more false positives.

Among the different models tested, RF consistently performed the best across all datasets. This indicates that RF's ability to handle high-dimensional data and its robustness to overfitting make it particularly suitable for this application.

Explainability in this model can be observed with SHAP as mentioned in the section 3.2.1. The figure below shows the importance of variables when including or not contrast information from the occlusion site contralateral patch.

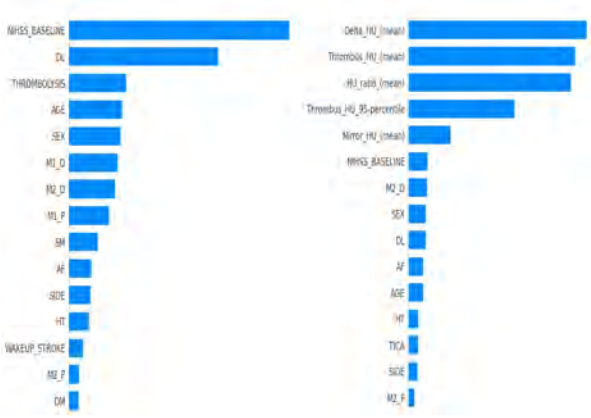


Figure 11: SHAP Summary plot showing the average predictor impact on model output in clinical data with and without the contrast information from the occlusion site contralateral patch (Siddiqui et al., 2023)

### *4.2. Imaging Data*

In this section, we present the best results obtained from the different experiments conducted on NCCT, CTA, and combined imaging modalities. These experiments utilized complete volumes with resampled voxel sizes as described in section 1.5. The subsequent experiments in this study are contingent upon the findings presented here. The results for this section are summarized in Table 4.

Table 3: Performance Metrics for Tabular Data

| Dataset | AUC | F1-score | Weighted Acc | Sensitivity | Specificity |
|---|---|---|---|---|---|
| Clinical | 0.59 ± 0.13 | 0.31 ± 0.07 | 0.70 ± 0.12 | 0.57 ± 0.26 | 0.72 ± 0.16 |
| Radiomics | 0.72 ± 0.06 | 0.43 ± 0.12 | 0.74 ± 0.12 | 0.70 ± 0.09 | 0.74 ± 0.14 |
| **Clinical + Radiomics** | **0.74 ± 0.13** | **0.45 ± 0.14** | **0.68 ± 0.18** | **0.88 ± 0.09** | **0.64 ± 0.21** |

An important consideration in these experiments is the comparison between using the whole volume versus skull-stripped volumes. Additionally, the methods used to address class imbalance during training — whether by adjusting the weights in the loss function or by upsampling/downsampling — play an important role in the outcomes. This section's findings will inform the best practices for preprocessing and balancing techniques in the subsequent multimodal integration experiments.

#### *4.2.1. Prior Information*

Prior information is presented with one case of loss as balancing mode and the others using sampling as the balancing method, as seen in table 5.

### *4.3. DAFT*

Table 6 presents the results of the best model of each balancing method. Table 7 shows the results of the best imaging mode model for each prior information type for DAFT with prior information and. Finally, daft models are tested using ensemble models, comparing prior and no prior approaches. Imaging technique is not shown as precisely, the different imaging techniques are part of the weights according to their performance.

## 5. Discussion

### *5.1. Tabular Data*

The performance of the clinical data alone indicates that while clinical variables are important for prediction, they may not provide sufficient accuracy when used in isolation. The relatively lower performance metrics suggest that additional data sources are needed to improve predictive capabilities.

Radiomics data showed a marked improvement over clinical data alone. This indicates that radiomic features capture important structural and textural information from the occlusion, which are highly relevant in this task. The use of scaled data followed by heavy RFE proved to be the most effective preprocessing strategy. However, it is important to note that obtaining radiomic features requires accurate thrombus segmentation. This step necessitates a robust segmentation method, as poor segmentation can lead to low-quality predictors, making this approach more impractical compared to methods that do not require such priors.

The combination of clinical and radiomics data yielded the best results, demonstrating that integrating multiple data sources can significantly enhance predictive performance. A key finding was that AF emerged as one of the top predictors according to SHAP analysis. AF is a risk factor linked to CE stroke and therefore has a high negative predictive value for the classification target at hand, which may explain this result. The scaling and RFE preprocessing strategy again proved to be the most effective, suggesting that careful preprocessing is crucial for maximizing the predictive power of combined datasets.

#### *5.1.1. Insights from Additional Experiments*

Additional experiments with the clinical data revealed that including contralateral patch information significantly improved prediction results. When this information was excluded, the top five predictors, according to SHAP analysis, were baseline NIHSS, dyslipidemia, IVT, age, and sex as seen in figure 11. These predictors remained top-performing even when additional radiologic predictors described in Siddiqui et al. (2023) were included, suggesting their strong and consistent relevance in predicting outcomes.

On the side of the radiomics dataset, the best results were obtained when data were scaled, and a rigorous RFE process was applied. This indicates that radiomics features are highly informative but require careful selection and preprocessing to enhance model performance.

### *5.2. Imaging Data*

The use of resampled volumes with loss-based class balancing methods generally resulted in lower performance metrics. For instance, EffB0 and DenseNet169 exhibited the lowest AUC scores (Table 4). Models demonstrated improved performance when employing sampling methods for class balancing. DenseNet169 achieved the highest AUC in these experiments, highlighting the effectiveness of this preprocessing and balancing strategy. This suggests that sampling methods are more suited for managing class imbalance in this context compared to adjusting loss weights.

Skull stripping with loss-based balancing did not yield significant improvements. Notably, DenseNet169 and EfficientNetB0 showed a zero F1-score and extremely high specificity with no sensitivity, indicating potential model collapse and inability to generalize from the data. On the other hand, skull stripping combined with sampling methods showed promising results. For example, ResNet34 achieved a notable AUC and a

Table 4: Performance Metrics for Different Models and Preprocessing Methods, Selecting the Best Imaging Mode for Each Model

| Model | Imaging Mode | AUC | F1-score | Weighted Acc | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| **Preprocessing: Resampled, Balancing Method: Loss** | | | | | | |
| ResNet34 | CTA | 0.61 ± 0.11 | 0.18 ± 0.11 | 0.53 ± 0.03 | 0.58 ± 0.31 | 0.52 ± 0.46 |
| DenseNet169 | BOTH | 0.47 ± 0.14 | 0.09 ± 0.16 | 0.53 ± 0.05 | 0.07 ± 0.01 | 0.97 ± 0.05 |
| EffB0 | BOTH | 0.61 ± 0.11 | 0.18 ± 0.11 | 0.53 ± 0.03 | 0.58 ± 0.31 | 0.52 ± 0.46 |
| **Preprocessing: Resampled, Balancing Method: Sampling** | | | | | | |
| ResNet34 | BOTH | 0.55 ± 0.16 | 0.30 ± 0.09 | 0.64 ± 0.27 | 0.58 ± 0.35 | 0.65 ± 0.36 |
| DenseNet169 | BOTH | 0.67 ± 0.04 | 0.34 ± 0.04 | 0.65 ± 0.03 | 0.67 ± 0.14 | 0.47 ± 0.24 |
| EffB0 | NCCT | 0.63 ± 0.09 | 0.28 ± 0.05 | 0.60 ± 0.08 | 0.33 ± 0.13 | 0.07 ± 0.06 |
| **Preprocessing: Skull Stripped, Balancing Method: Loss** | | | | | | |
| ResNet34 | CTA | 0.52 ± 0.09 | 0.00 ± 0.00 | 0.50 ± 0.00 | 0.00 ± 0.00 | 1.00 ± 0.00 |
| DenseNet169 | NCCT | 0.66 ± 0.05 | 0.00 ± 0.00 | 0.50 ± 0.00 | 0.00 ± 0.00 | 1.00 ± 0.00 |
| EffB0 | BOTH | 0.55 ± 0.06 | 0.13 ± 0.14 | 0.53 ± 0.06 | 0.63 ± 0.31 | 0.61 ± 0.42 |
| **Preprocessing: Skull Stripped, Balancing Method: Sampling** | | | | | | |
| ResNet34 | NCCT | 0.54 ± 0.06 | 0.32 ± 0.03 | 0.78 ± 0.05 | 0.55 ± 0.12 | 0.84 ± 0.06 |
| DenseNet169 | NCCT | 0.57 ± 0.13 | 0.36 ± 0.14 | 0.62 ± 0.08 | 0.56 ± 0.30 | 0.41 ± 0.39 |
| EffB0 | BOTH | 0.59 ± 0.17 | 0.32 ± 0.12 | 0.60 ± 0.11 | 0.42 ± 0.29 | 0.23 ± 0.30 |

Table 5: Performance Metrics for Using Prior Information

| Model | Imaging Mode | AUC | F1-score | Weighted Acc | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| **Balancing Method: Loss** | | | | | | |
| EffB0 | Vessel (BOTH) | 0.58 ± 0.12 | 0.23 ± 0.14 | 0.59 ± 0.09 | 0.61 ± 0.23 | 0.45 ± 0.36 |
| **Balancing Method: Sampling** | | | | | | |
| EffB0 | Vessel (BOTH) | 0.62 ± 0.14 | 0.30 ± 0.07 | 0.63 ± 0.07 | 0.48 ± 0.15 | 0.20 ± 0.13 |
| EffB0 | Thrombus (BOTH) | 0.57 ± 0.11 | 0.30 ± 0.04 | 0.62 ± 0.04 | 0.59 ± 0.04 | 0.35 ± 0.19 |
| EffB0 | Both Prior (BOTH) | 0.62 ± 0.05 | 0.27 ± 0.05 | 0.61 ± 0.06 | 0.60 ± 0.24 | 0.45 ± 0.38 |

higher F1-score compared to other preprocessing methods. This indicates that removing extraneous information and using sampling methods can improve model performance by focusing on the most relevant regions in the imaging data.

Given the heavily imbalanced dataset, addressing class imbalance is critical. The experiments demonstrated that sampling methods consistently provided better performance metrics compared to adjusting the weights in the loss function. The sampling approach effectively balanced the class distribution during training, leading to more stable and reliable model performance. Therefore, sampling was chosen as the primary class balancing technique for subsequent experiments.

In the comparison between using the whole volume versus skull-stripped volumes, the results indicated that complete volumes yielded better performance metrics overall. Skull stripping did not significantly enhance model performance and, in some cases, led to poorer generalization. The complete volume approach ensures that all relevant anatomical information is preserved, which may include subtle features necessary for accurate predictions. For instance, skull stripping could unintentionally remove potential atheromatosis in the intracranial part of the internal carotid artery, a feature that could be predictive of ICAD. Consequently, we decided to proceed with complete volumes for all further experiments, answering the question proposed in section 1.5.

Among the different models evaluated, EffB0 consistently showed competitive performance. Its advanced architecture, which uses a compound scaling method to optimize depth, width, and resolution, proved effective in handling the complexity of the imaging data. Additionally, EfficientNet's efficiency in terms of computational resources makes it a practical choice for extensive experimentation. Therefore, it was decided to use EfficientNet exclusively for all subsequent analyses.

#### *5.2.1. Imaging Data with Prior Information*

The performance metrics for experiments using prior information and different balancing methods are presented in Table 5.

Incorporating vascular segmentation as prior information with loss-based balancing showed some improvement, but it was not substantial. On the other hand, using thrombus information as prior yielded moderate effectiveness with good AUC and F1-scores. Using both vessel and thrombus information together showed similar AUC to using vessel information alone but with slightly lower F1-scores and weighted accuracy.

These findings suggest that prior information, such as vessel and thrombus segmentations, can aid in model regularization and performance enhancement. This

Table 6: Performance Metrics for DAFT

| Model | Imaging Mode | AUC | F1-score | Weighted Acc | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| **Balancing Method: Loss** | | | | | | |
| DaftEffB0 | BOTH | 0.53 ± 0.09 | 0.25 ± 0.02 | 0.55 ± 0.05 | 0.42 ± 0.29 | 0.27 ± 0.29 |
| **Balancing Method: Sampling** | | | | | | |
| DaftEffB0 | NCCT | 0.67 ± 0.02 | 0.36 ± 0.03 | 0.69 ± 0.04 | 0.68 ± 0.05 | 0.43 ± 0.08 |

Table 7: Performance Metrics for DAFT with Prior Information using EfficientNet and Balancing: Sampling

| Model | Imaging Mode | AUC | F1-score | Weighted Acc | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| **Prior: Vessel** | | | | | | |
| DaftEffB0 | NCCT | 0.62 ± 0.10 | 0.36 ± 0.09 | 0.68 ± 0.08 | 0.64 ± 0.12 | 0.38 ± 0.14 |
| **Prior: Thrombus** | | | | | | |
| DaftEffB0 | CTA | 0.65 ± 0.12 | 0.34 ± 0.04 | 0.68 ± 0.05 | 0.53 ± 0.14 | 0.24 ± 0.19 |
| **Prior: Both Vessel and Thrombus** | | | | | | |
| DaftEffB0 | BOTH | 0.64 ± 0.10 | 0.37 ± 0.11 | 0.68 ± 0.07 | 0.67 ± 0.19 | 0.47 ± 0.24 |

might be due to these segmentations providing additional relevant context, which helps the model distinguish between different regions of interest and pathological features more effectively. By focusing on specific regions of interest, the model can learn more meaningful patterns that contribute to better prediction outcomes.

However, adding prior information in this manner to the imaging-only experiments did not yield improvements over methods based on radiomics. The complexity of the data, high dimensionality, and relatively small dataset size may explain this effect. These factors likely contribute to the model's ability to generalize from the training data, impacting the overall performance.

Sampling methods were more effective in dealing with class imbalance compared to loss-based balancing, as evidenced by the higher AUC and F1-scores achieved with sampling methods. Sampling ensures a balanced class distribution during training, leading to better generalization and performance on the minority class. While using both vessel and thrombus information did not significantly outperform using vessel information alone, this may indicate the particular relevance of vessel information for this prediction task.

### *5.3. DAFT*

The performance metrics for the DAFT model, summarized in Table 6, provide insights that highlight the benefits of integrating imaging data with clinical information.

The DaftEffB0 model using both NCCT imaging and tabular data demonstrated that incorporating clinical data significantly enhances model performance. While the loss-based balancing method indicated relatively lower effectiveness, it is notable that our modified DAFT version achieved its best performance with this method, reflecting the potential of integrating diverse data sources. However, overall, sampling methods consistently showed superior results, suggesting their greater suitability for managing class imbalance in this context, as observed with image-based models.

One important aspect pointing to the improvement made by the multimodal approach is comparing the results of the just clinical approach seen in Table 3, just NCCT imaging with sampling in Table 4, and the result of using both data sources in DAFT. Clear improvements in AUC and F1-score indicate better discrimination ability and more accurate predictions. Weighted accuracy remained similar between tabular and multimodal methods, suggesting consistent overall performance. An improvement in sensitivity at the cost of reduced specificity means the model becomes better at identifying true positive cases but may generate more false positives.

#### *5.3.1. Prior DAFT*

The results here indicate that the integration of prior information with the DAFT model leads to improvements in several metrics.

Using vessel as prior, the best model using DAFT is consistent with the best model without DAFT, but there are improvements in F1-score, AUC and weighted accuracy compared to models without clinical data, at the cost of losing specificity. In the case of thrombus as prior where there are also higher AUC, F1-score and weighted accuracy, sensitivity and specificity have been affected, making the new best model more balanced in handling true positives and true negatives but at the cost of more variability in detecting true positives and true negatives.

Using both vessel and thrombus as prior information, DAFT with combined imaging modes improves all metrics. This highlights the benefit of combining multiple sources of prior information for better model performance despite the consideration of having all data beforehand in a clinical context.

Overall, the DAFT model incorporating prior information outperformed the models using prior informa-

Table 8: Performance Metrics for Ensemble models using DAFT with Resampled Prior Information

| Model | AUC | F1-score | Weighted Acc | Sensitivity | Specificity |
|---|---|---|---|---|---|
| **DAFT No Prior** | 0.64 ± 0.07 | 0.34 ± 0.07 | 0.48 ± 0.20 | 0.66 ± 0.15 | 0.61 ± 0.27 |
| **DAFT Prior Vessel** | **0.74 ± 0.12** | **0.42± 0.15** | **0.54± 0.06** | **0.76± 0.18** | **0.69± 0.08** |
| **DAFT Prior Thrombus** | 0.68 ± 0.04 | 0.35± 0.06 | 0.43 ± 0.18 | 0.81 ± 0.23 | 0.55 ± 0.25 |
| **DAFT Prior Both** | 0.71 ± 0.07 | 0.39 ± 0.08 | 0.52 ± 0.11 | 0.64 ± 0.07 | 0.68 ± 0.14 |

tion without clinical data, confirming the importance of integrating diverse data sources. However, it is notable that not always the best performing setup for each prior method remains consistent with and without DAFT, as it happens in vessel, suggesting that while DAFT adds value, the fundamental importance of certain prior information remains unchanged.

#### *5.3.2. Ensemble DAFT*

The results from Table 8 shows that the ensemble models incorporating prior information generally improved the performance compared to the DAFT model without prior information.

Firstly, it is evident that the ensemble models using prior information generally show improved performance metrics compared to the best individual models. For instance, the Prior Vessel model achieves a great AUC, which is significantly higher than the best individual model, and it is the best model of the study. This improvement in AUC, along with the F1-score and weighted accuracy, highlights the effectiveness of the ensemble approach in enhancing model performance by using multiple prior information sources. The higher sensitivity in the ensemble models indicates better identification of positive cases, showing that the ensemble method helps in reducing false negatives.

Secondly, the DAFT Resampled Prior Both model, which uses both vessel and thrombus information, shows a substantial improvement in AUC compared to the best individual model using both priors. The increase in F1-score and weighted accuracy in the ensemble models also suggests that integrating diverse data sources through ensembling helps in achieving a balanced performance across different metrics.

However, it is important to note that while the ensemble models generally perform better, there are trade-offs. The Prior Thrombus model, for instance, shows an increase in sensitivity but a decrease in specificity, indicating a higher rate of false positives. This could be due to the ensemble model's increased focus on identifying positive cases, potentially leading to overfitting on the training data. The confusion matrices further support this, showing that ensemble models have a more balanced distribution of true positives and true negatives, but at the cost of some increase in false positives.

Interestingly, the No Prior model did not benefit as much from the ensemble approach, as indicated by an AUC, which is lower compared to the best individual DAFT model without prior information, that was the best model until this step. This suggests that the ensemble method may not always improve performance and can sometimes lead to reduced effectiveness, especially when the base models do not capture sufficient complementary information. The decrease in specificity and the relatively unchanged sensitivity and F1-score highlight the limitations of ensembling in cases where prior information is not used or where the models that will be part of the ensemble are either not good performing, nor outstanding in at least classifying true positives or true negatives.

Overall, these findings suggest that ensemble methods can effectively improve model performance by integrating diverse sources of information. However, careful tuning and validation are required to manage the trade-offs between sensitivity and specificity, ensuring the model generalizes well to new data. The results also emphasize that ensembling relies on the assembling method and that the quality and nature of the base models play a crucial role in the success of ensemble approaches.

### *5.4. Practical Implications*

The research findings emphasize the importance of integrating prior information, such as thrombus and vessel segmentations, to significantly enhance the predictive performance of models for ICAD. These enhancements are evident in improved AUC, F1-scores, and weighted accuracy metrics, indicating that prior information helps models focus on critical regions, thereby improving class differentiation.

However, the practical implementation of such models in clinical settings demands robust and accurate segmentation methods. High-quality imaging and precise delineation of thrombus and vessel regions are essential to ensure reliable data input for the models. For instance, advanced pre-operative imaging techniques like CTA are beneficial for effective vessel segmentation. Thus, ensuring the robustness of segmentation methods is vital for clinical application.

Implementing these models also requires substantial computational resources, particularly in hospital settings where quick processing of large volumes of imaging data is crucial. High-performance computing systems must be available to minimize segmentation and classification time, aligning with the workflow of

clinicians and allowing timely decision-making during emergency procedures.

Despite the promising results, several limitations must be acknowledged. The dataset size was limited, which may affect the robustness and generalizability of the findings. The definition of the class of interest was not highly specific, potentially leading to data heterogeneity. The study was conducted in a single center with retrospective data, which may limit the applicability of the results to other settings. Additionally, the time available for conducting experiments was limited, restricting the extent of model optimization and validation.

The primary goal of these predictive models is to provide clinicians with preoperative insights about the potential presence of ICAD before performing EVT. Accurate predictions can guide the selection of appropriate treatment strategies and improve patient outcomes. Future clinical trials should focus on datasets specifically tailored to ICAD, exploring the efficacy of different rescue treatments and considering the total inference time of models to ensure real-time applicability.

Beyond identifying patients with ICAD or predicting unsuccessful outcomes with conventional treatments, a model with strong discriminative capability could have broader implications. It could aid in personalizing treatment plans, selecting alternative therapies such as stenting or angioplasty, and improving the overall management of stroke patients.

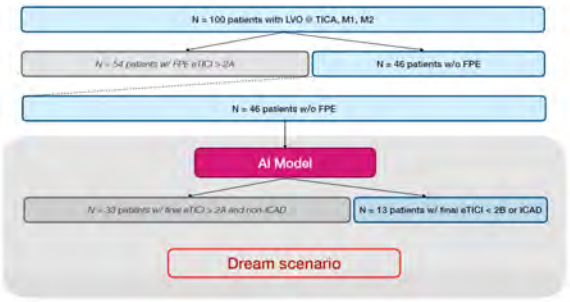



Figure 12: Diagram of a potential clinical trial using an AI model as a patient selection tool. Negative Group: Gray after AI model, would go through conventional treatment and positive group would go through, for example, one EVT pass + stenting.

Good discriminators can also be helpful in the randomization of clinical trials, ensuring that the outcome of conventional treatment versus tested rescue treatments is more reliable. Figure 12 shows an diagram of how a validated AI model could be integrated for patient selection in a randomized control trial. By accurately identifying ICAD-affected individuals, such models can reduce the hazard of multiple EVT passes or inappropriate stenting in CE patients, thus improving treatment adequacy on a selected population.

Additionally, integrating medical reports to curate tabular data and utilizing large language models (LLMs) to generate informative embeddings can enhance multimodal integration, potentially improving the accuracy and relevance of predictive models in clinical settings.

Finally is worth to mention that our best ensemble model with multimodal data achieved better results than the random bootstrapped classifier, which no one could beat in the multimodal integration part of the IACTA-EST 2023 challenge (AUC: 0.27-0.73) mentioned in section 2.4. This promising outcome indicates the potential of our approach to advance the field and improve clinical practice in the ICAD-LVO field.

## 6. Conclusions

This thesis explored the integration of multimodal data to predict failed recanalization in patients undergoing EVT for large vessel occlusion (LVO) strokes. By combining clinical data, radiomics, and imaging modalities, we aimed to enhance prediction accuracy and provide valuable insights for clinical decision-making.

Our findings show that integrating clinical and radiomics data significantly improves predictive performance compared to using clinical data alone. The structural and textural information captured by radiomics is essential in understanding thrombus pathology and treatment response. Additionally, the Dynamic DAFT model further improved performance metrics by combining diverse data sources, leading to better discrimination and more accurate predictions.

Incorporating prior information, such as vessel and thrombus segmentations, into the DAFT model provided additional context, improving model regularization and performance. This approach allowed the model to focus on specific regions of interest, leading to better prediction outcomes.

Despite these promising results, several limitations must be acknowledged. The small dataset size may limit the robustness and generalizability of the findings. Larger, multi-center datasets are needed to validate the results. The study was conducted in a single center with retrospective data analysis, which may limit the applicability of the results to other settings. The time constraints also restricted the extent of model optimization and validation.

Our study demonstrates the potential of using advanced machine learning models to provide accurate preoperative predictions of failed recanalization in stroke patients. Future work should aim to validate these findings in larger cohorts, refine the models for broader clinical application, and ultimately contribute to more effective stroke management strategies.

## Acknowledgments

I would firstly like to express my gratitude to God Almighty. To Pere Canals for his invaluable insights in the work done. To Marc Ribo for the welcoming support

in the Stroke Research Unit at the Vall d'Hebron Hospital Campus and research institute and to Alvaro Garcia-Tornel for his invaluable clinical-medical support.

Nothing but my deep gratitude to the MAIA consortium and to Santander Open Academy. For the friends that have supported me and the friends I have made here. Thanks for the appreciated support from mi family.

## References

Amukotuwa, S., Straka, M., Aksoy, D., Fischbein, N., Desmond, P., Albers, G., Bammer, R., 2019. Cerebral blood flow predicts the infarct core: New insights from contemporaneous diffusion and perfusion imaging. Stroke 50. doi:10.1161/STROKEAHA.119.026640.

Banerjee, C., Chimowitz, M.I., 2017. Stroke caused by atherosclerosis of the major intracranial arteries. Circulation research 2, 502–513. doi:10.1161/CIRCRESAHA.116.308441.

Bang, O.Y., Kim, B.M., Seo, W., Jeon, P., 2010. Endovascular therapy for acute ischemic stroke of intracranial atherosclerotic origin—neuroimaging perspectives. Frontiers in Neurology 10. doi:https://doi.org/10.3389/fneur.2019.00269.

Beaman, C., Yaghi, S., Liebeskind, D.S., 2022. A decade on: The evolving renaissance in intracranial atherosclerotic disease. Stroke: Vascular and Interventional Neurology 2. doi:10.1161/SVIN.122.000497.

Bento, M., Souto-Maior-Neto, L., Salluzzi, M., Zhang, Y., R., F., 2018. Feature extraction using convolutional networks for identifying carotid artery atherosclerosis patients in a heterogeneous brain mr dataset. Proceedings of Joint Annual Meeting of International Society for Magnetic Resonance in Medicine .

Bento, M., Souza, R., Salluzzi, M., Rittner, L., Zhang, Y., R., F., 2019. Automatic identification of atherosclerosis subjects in a heterogeneous mr brain imaging data set. Magnetic Resonance Imaging 62, 18–27. doi:https://doi.org/10.1016/j.mri.2019.06.007.

Cai, Y., Gu, Y., Wang, Y., Wang, P., Zhang, L., Liu, C., Chu, J., Li, H., Lu, Z., Zhou, Y., Liu, H., 2022. A clinical prediction model for patients with acute large vessel occlusion due to underlying intracranial atherosclerotic stenosis. Clinical Neuroradiology 33, 519–528. doi:10.1007/s00062-022-01241-3.

Canals, P., Balocco, S., Diaz, O., Li, J., Garcia-Tornel, A., Tomasello, A., Olive-Gadea, M., Ribo, M., 2023. A fully automatic method for vascular tortuosity feature extraction in the supra-aortic region: unraveling possibilities in stroke treatment planning. Computerized Medical Imaging and Graphics 104. doi:https://doi.org/10.1016/j.compmedimag.2022.102170.

Cardoso, J.M., 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., 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., A., D., Maier-Hein, L., Jaeger, P.F., Baumgartner, M., J., K.C., Flores, M., Kirby, J., D., C.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. doi:https://doi.org/10.48550/arXiv.2211.02701.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA. pp. 785–794. doi:10.1145/2939672.2939785.

Chen, W., Liu, J., Yang, L., Sun, H., Yang, S., Wang, M., Qin, W., Wang, Y., Wang, X., Hu, W., 2023. Development and internal-external validation of the athe scale. Journal of Neurosurgery doi:10.3171/2023.10.JNS232084.

Coutts, S.B., 2017. Thrombus composition and efficacy of thrombolysis and thrombectomy in acute ischemic stroke. CONTINUUM: Lifelong Learning in Neurology 23, 82–92. doi:10.1212/CON.0000000000000424.

Cui, L., Fan, Z., Yang, Y., Liu, R., Wang, D., Feng, Y., Lu, J., Fan, Y., 2022. Deep learning in ischemic stroke imaging analysis: A comprehensive review. BioMed Research International doi:https://doi.org/10.1155/2022/2456550.

Dietterich, T.G., 2000. Ensemble methods in machine learning. Multiple Classifier Systems , 1–15doi:10.1007/3-540-45014-9_1.

Dundamadappa, S., Iyer, K., Agrawal, A., Choi, D.J., 2021. Multiphase ct angiography: A useful technique in acute stroke imaging—collaterals and beyond. AJNR Am J Neuroradiol 42. URL: https://www.ajnr.org.

Fischer, U., Kaesmacher, J., Plattner, P., Butikofer, L., Mordasini, P., Deppeler, S., Cognard, C., Pereira, V.M., Siddiqui, A., T., F.M., Furlar, A.J., Chapot, R., Strbian, D., Wiesmann, M., Bressan, J., Lerch, S., Liebeskind, D.S., Saver, J.L., Gralla, J., 2022. Swift direct: Solitaire™ with the intention for thrombectomy plus intravenous t-pa versus direct solitaire™ stent-retriever thrombectomy in acute anterior circulation stroke: Methodology of a randomized, controlled, multicentre study. International Journal of Stroke 17, 698–705. doi:10.1177/17474930211048768.

Goyal, M., Menon, B.K., van Zwam, W.H., Dippel, D.W.J., Mitchell, P.J., Demchuk, A.M., Dávalos, A., Majoie, C.B.L.M., Saver, J.L., Levy, E.I., Campbell, B.C.V., Hacke, W., White, P.M., Pereira, V.M., Köhrmann, T., Lopes, D.K., Fiehler, E.J.F., Bonafe, A., Diener, H.C., Vermeij, J.A.F., Hill, A.M., van der Worp, H.B., Roos, Y.B.W.E.M., Audebert, H.J., Fischer, U., Albers, G.W., Schellinger, P.D., Rudd, A.S., Markus, H.S., Marquering, J.G., Majoie, E.M.M., Berkhemer, O.A., Majoie, C.B.L., 2016. Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. The Lancet 387, 1723–1731. doi:10.1016/S0140-6736(16)00163-X.

Grotta, J.C., 2023. Intravenous thrombolysis for acute ischemic stroke. CONTINUUM: Lifelong Learning in Neurology 29, 425–442. doi:10.1212/CON.0000000000001207.

Guyon, E., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Machine Learning 46, 389–422.

Haussen, D.C., Bouslama, M., Dehkharghani, S., Grossberg, J.A., Bianchi, N., Bowen, M., Frankel, M.R., Nogueira, R.G., 2018. Automated ct perfusion prediction of large vessel acute stroke from intracranial atherosclerotic disease. Interventional Neurology 7, 334–340. doi:10.1159/000487335.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. doi:https://doi.org/10.48550/arXiv.1512.03385.

Huang, G., Liu, Z., Van der Maaten, L., Weinberger, K.Q., 2018. Densely connected convolutional networks. doi:https://doi.org/10.48550/arXiv.1608.06993.

(IHME), I.f.H.M., Evaluation., 2021. Global burden of disease study 2021 (gbd 2021) results. URL: https://vizhub.healthdata.org/gbd-results/.

Jolugbo, P., Ariëns, R.A., 2021. Diagnosis and management of transient ischemic attack. Stroke 52, 1131–1142. doi:10.1161/STROKEAHA.120.032810.

Lal-trehan, U., Giancardo, L., 2021. Optimizing stroke segmentation using acute brain cta.

Li, H., Ma, H.Y., Zhang, L., Liu, P., Zhang, Y.X., Zhang, X.X., Li, Z.F., Peng-Fei, X., Yong-Wei, Z., Li, Q., Peng-Fei, Y., Liu, J.M., 2022. Early diagnosis of intracranial atherosclerotic large vascular occlusion: A prediction model based on direct-mt data. Frontiers in Neurology 13. doi:10.3389/fneur.2022.1026815.

Liao, G., Zhang, Z., Tung, T.H., He, Y., Hu, L., Zhang, X., Chen, H., Huang, J., Du, W., Li, C., Yang, Z., Cai, Y., Liang, H., 2022. A simple score to predict atherosclerotic or embolic intracranial large-vessel occlusion stroke before endovascular treatment. Journal of Neurosurgery 137, 1501–1508. doi:10.3171/2022.1.JNS212924.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Ad-

vances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 4765–4774.

Lyndon, D., Van De Broek, M., Niu, B., Yip, S., Rohr, A., Settecase, F., 2021. Hypoperfusion intensity ratio correlates with cta collateral status in large-vessel occlusion acute ischemic stroke. American Journal of Neuroradiology 42, 1380–1386. doi:10.3174/ajnr.A7181.

Maguida, G., Shuaib, A., 2023. Collateral circulation in ischemic stroke: An updated review. Journal of Stroke 25, 179–198. doi:10.5853/jos.2022.02936.

Maida, C.D., Norrito, M.L., Daidone, M., Tuttolomondo, A., Pinto, A., 2020. Neuroinflammatory mechanisms in ischemic stroke: Focus on cardioembolic stroke, background, and therapeutic approaches. International Journal of Molecular Sciences 21. doi:https://doi.org/10.3390/ijms21186454.

Meschia, J.F., 2023. Diagnostic evaluation of stroke etiology. CONTINUUM: Lifelong Learning in Neurology 29, 412–424. doi:10.1212/con.0000000000001206.

Montavon, G., Orr, G.B., Müller, K.R. (Eds.), 2012. Neural Networks: Tricks of the Trade, Second Edition. volume 7700 of *Lecture Notes in Computer Science*. Springer, Heidelberg. doi:10.1007/978-3-642-35289-8.

Patel, A., 2023. Automated Image Analysis of Cranial Non-Contrast CT. Ph.D. thesis. Radboud University Nijmegen. Nijmegen, The Netherlands. doi:10.6100/IR292990. available at `https://repository.ubn.ru.nl/handle/2066/292990`.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Pexman, J.H., Barber, P.A., Hill, M.D., Sevick, R.J., Demchuk, A.M., Hudon, M.E., Hu, W.Y., Buchan, A.M., 2001. Use of the alberta stroke program early ct score (aspects) for assessing ct scans in patients with acute stroke. AJNR Am J Neuroradiol 22, 1534–1542.

Phan, K., Dmytriw, A.A., Maingard, J., Asadi, H., Griessenauer, C.J., Ng, W., Kewagamang, K., Mobbs, R.J., Moore, J.M., Ogilvy, C.S., et al., 2017. Endovascular thrombectomy alone versus combined with intravenous thrombolysis. World Neurosurgery 108. doi:10.1016/j.wneu.2017.08.040.

Polsterl, S., Wolf, T.N., Wachinger, C., 2021. Combining 3D Image and Tabular Data via the Dynamic Affine Feature Map Transform. Springer International Publishing. p. 688–698. doi:10.1007/978-3-030-87240-3_66.

Psychogios, M., Brehm, A., López-Cancio, E., et al., 2022. European stroke organisation guidelines on treatment of patients with intracranial atherosclerotic disease. European Stroke Journal 7, XLII–LXXX. doi:10.1177/23969873221099715.

Qiu, J., Tan, G., Lin, Y., Guan, J., Dai, Z., Wang, F., Zhuang, C., Wilman, A., Huang, H., Cao, Z., Tang, Y., Jia, Y., Li, Y., Zhou, T., Wu, R., 2022. Automated detection of intracranial artery stenosis and occlusion in magnetic resonance angiography: A preliminary study based on deep learning. Magnetic Resonance Imaging 94, 105–111. doi:10.1016/j.mri.2022.09.006.

Rodrigo-Gisbert, M., Garcia-Tornel, A., Requena, M., Vielba-Gomez, I., Bashir, S., Rubiera, M., Lascuevas, M.d.D., Olivé-Gadea, M., Piñana, C., Rizzo, F., Muchada, M., Rodriguez-Villatoro, N., Rodriguez-Luna, D., Juega, J., Pagola, J., Hernandez, D., Molina, C.A., Terceño, M., Tomasello, A., Ribo, M., 2024. Clinico-radiological features of intracranial atherosclerosis-related large vessel occlusion prior to endovascular treatment. Scientific Reports 14. doi:10.1038/s41598-024-53354-z.

Rodrigo-Gisbert, M., Requena, M., Rubiera, M., Khalife, J., Lozano, P., De Dios Lascuevas, M., Garcia-Tornel, A., Olive-Gadea, M., Piñana, C., Rizzo, F., Boned, S., Muchada, M., Rodriguez-Villatoro, N., Rodriguez-Luna, D., Juega, J., Pagola, J., Hernandez, D., Molina, C.A., Tomasello, A., Ribo, M., 2023. Intracranial artery calcifications profile as a predictor of recanalization failure in endovascular stroke treatment. Stroke 2, 430–438. doi:10.1161/STROKEAHA.122.041257.

Rodriguez-Calienes, A., Siddiqui, F., Vivanco-Suarez, J., Shogren, S., Galecio-Castillo, M., Dibas, M., Pandey, A., Ribo, M., Ortega-Gutierrez, S., 2024. Unmasking the imitators: Challenges in identifying intracranial atherosclerosis-related large vessel occlusion mimics during mechanical thrombectomy. Stroke: Vascular and Interventional Neurology 0. doi:10.1161/SVIN.123.001303.

Sanchez, S., Mossa-Basha, M., Anagnostakou, V., Liebeskind, D., Samaniego, E., 2024. Comprehensive imaging analysis of intracranial atherosclerosis. J NeuroIntervent Surg doi:10.1136/jnis-2023-020622.

Saver, J.L., 2006. Time is brain–quantified. Stroke: Vascular and Interventional Neurology 1. doi:10.1161/01.STR.0000196957.55928.ab.

Sheth, S.A., 2023. Mechanical thrombectomy for acute ischemic stroke. CONTINUUM: Lifelong Learning in Neurology 29, 443–461. doi:10.1212/con.0000000000001243.

Shulman, J.G., Abdalkader, M., 2023. Imaging of central nervous system ischemia. CONTINUUM: Lifelong Learning in Neurology 29, 54–72. doi:10.1212/con.0000000000001185.

Siddiqui, F.M., Fletcher, J.J., Barnes, A.V., Henry, A.N., Elias, A.E., Rajah, G., Carroll, A., Dandapat, S., Ume, K.L., Farooqi, M., Rodriguez-Calienes, A., S., P.A., Ortega-Gutierrez-Santiago, 2023. External validation of atherosclerotic neuroimaging biomarkers in emergent large-vessel occlusion. Stroke: Vascular and Interventional Neurology 3. doi:10.1161/SVIN.123.000850.

Tan, M., Le, Q.V., 2020. Efficientnet: Rethinking model scaling for convolutional neural networks. doi:https://doi.org/10.48550/arXiv.1905.11946.

van Voorst, H., Bruggeman, A.A.E., Yang, W., Andriessen, J., Welberg, E., Dutra, B.G., Konduri, P.R., Arrarte Terreros, N., Hoving, J.W., Tolhuisen, M.L., Kappelhof, M., Brouwer, J., Boodt, N., van Kranendonk, K.R., Koopman, M.S., Hund, H.M., Krietemeijer, M., van Zwam, W.H., van Beusekom, H.M.M., van der Lugt, A., Emmer, B.J., Marquering, H.A., Roos, Y.B.W.E.M., Caan, M.W.A., Majoie, C.B.L.M., 2023. Thrombus radiomics in patients with anterior circulation acute ischemic stroke undergoing endovascular treatment. Journal of Neurointerventional Surgery 15, e79–e85. URL: `https://pubmed.ncbi.nlm.nih.gov/35882552`, doi:10.1136/jnis-2022-019085.

Wasserthal, J., Breit, H., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M., 2023. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence 5. doi:10.1148/ryai.230024.

Wing, S.C., Markus, H.S., 2019. Interpreting ct perfusion in stroke. Practical Neurology 19, 136–142. doi:10.1136/practneurol-2018-001917.

Yeo, T., Wallace, R., Partovi, S., Institute, B.N., 2001. Ct angiography and stroke. Barrow Quarterly 17. URL: `https://www.barrowneuro.org`.

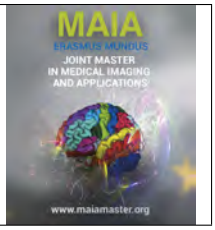
Yusuying, S., Lu, Y., Zhang, S., Wang, J., Chen, J., Wang, D., Lu, J., Qi, P., 2023. Ct-based thrombus radiomics nomogram for predicting secondary embolization during mechanical thrombectomy for large vessel occlusion. Frontiers in Neurology 14. doi:10.3389/fneur.2023.1152730.

Zaidat, O.O., Castonguay, A.C., Linfante, I., Gupta, R., Martin, C.O., Holloway, W.E., Mueller-Kronast, N., English, J.D., Dabus, G., Malisch, T.W., Marden, F.A., Bozorgchami, H., Xavier, A., Rai, A.T., T., F.M., Badruddin, A., Nguyen, T.N., Taqi, M.A., Abraham, M.G., G., Y.A., Janardhan, V., Shaltoni, H., Novakovic, R., Abou-Chebl, A., Chen, P.R., Britz, G.W., Sun, C.J., Bansal, V., Kaushal, R., Nanda, A., Nogueira, R.G., 2018. First pass effect: a new measure for stroke thrombectomy devices. Stroke 49. doi:10.1161/STROKEAHA.117.020315.

Zha, M., Wu, M., Huang, X., Xiaohao, Z., Huang, K., Qingwen, Y., Cai, H., Ji, Y., Lv, Q., Yang, D., Dai, Q., Liu, R., Liu, X., 2021. A pre-interventional scale to predict in situ atherosclerotic thrombosis in acute vertebrobasilar artery occlusion patients. Frontiers in Neurology 12. doi:10.3389/fneur.2021.648081.

Medical Imaging and Applications

Master Thesis, June 2024




# MMG-CLIP: Automated Mammography Reporting through Image-to-Text Translation

Abdelrahman Habib[a], Santiago Pires[b], Jaap Kroes[b]

[a]*Universitat de Girona, Spain; University of Bourgogne, France; Università degli studi di Cassino e del Lazio Meridionale, Italy*
[b]*ScreenPoint Medical, Nijmegen, Netherlands*

**Abstract**

Recently medical image-text datasets have become increasingly important in the development of deep learning applications, including automated radiology report generation models. Generating clinically valid radiology reports comes along with challenges, such as bridging the gap between interpreting medical images and accurately conveying the findings into radiology text reports. In this work, we tackle the task of automated mammography report generation following Breast Imaging Reporting & Data System (BI-RADS) guidelines. We utilize an image-label and exam-reports datasets, along with text prompting techniques, to generate a well-structured text report that supports training. Our proposed framework allows the usage of up to four image views within the exam, leveraging different information that can be captured from all exam views related to the radiology report. Our model demonstrated high performance in supervised and zero-shot classification settings when evaluated on multiple downstream tasks, enabling report generation as a series of zero-shot classification tasks.

*Keywords:* Mammography 2D X-ray, BI-RADS Report Generation, Contrastive Learning, Natural Language Processing

## 1. Introduction

Medical images from different modalities such as Mammography X-ray , Magnetic Resonance Imaging (MRI), and Computed Tomography (CT) are widely used to evaluate, monitor, and diagnose several medical conditions in clinical practice. Mammography X-ray is a universally accepted method for breast cancer detection as it is relatively in-expensive, repeatable, and widely available (Fishman and Rehani, 2021). Several applications demonstrated the effectiveness of deep-learning based models on solving tasks related to breast cancer detection in mammography, such in discrimination of microcalcifications (Wang et al., 2016), microcalcifications detection (Pesapane et al., 2023), breast cancer risk discrimination (Yala et al., 2019), and breast cancer image segmentation (Salama and Aly, 2021), and many others (Kallenberg et al., 2016; Mohamed et al., 2018; Ribli et al., 2018).

Although deep-learning models, such as convolutional neural networks (CNNs) by He et al. (2016); Krizhevsky et al. (2012); Simonyan and Zisserman (2014) have been widely applied for various artificial intelligence (AI) tasks in recent years (Han et al., 2021), and has been actively used for the purpose of medical image analysis (Anwar et al., 2018), the small size of annotated and publicly available medical datasets remains a major bottleneck in this area for developing computer-aided detection/diagnosis (CAD) tools. Unlike publicly available computer vision dataset that are available in large-scale, such as ImageNet (Deng et al., 2009) or OpenImages (Kuznetsova et al., 2020), publicly available medical datasets are much smaller in magnitude (Xie et al., 2021). This introduces challenges in training deep-learning models for medical purposes as the availability of high-quality clinical annotations is time-consuming an costly (You et al., 2023), and obtaining labels for medical images is very resource-intensive as it relies on domain experts (Karimi et al., 2020). Therefore, building effective medical imaging models is limited by the lack of large-scale annotated medical dataset.

Recently, Contrastive Language-Image Pre-training

(CLIP) as in the work of Radford et al. (2021), has achieved considerable success in computer vision and natural language processing domains, by allowing joint-training of image and text representation on large-scale image-text pairs (Wang et al., 2022), enabling zero-shot transfer of the model to downstream tasks. As shown by Radford et al. (2021), zero-shot CLIP models are much more robust than equivalently accuracy supervised ImageNet models. In another work, ALIGN by Jia et al. (2021) similarly to CLIP trains dual-encoder architecture to learn the alignment of visual and language representations of image and text pairs using contrastive loss by leveraging noisy dataset of over one billion image alt-text pairs. Both ALIGN and CLIP shows great robustness on classification tasks with different image distributions (Jia et al., 2021).

Considering CLIP, adopting such large vision-text pre-training models to the medical domain is a non-trivial task due to CLIP's data-hungry nature that was trained on 400 million (image, text) pairs collected from the internet (Wang et al., 2022). In that context, the natural solution of limited annotated medical dataset is to leverage the corresponding medical reports that contain detailed description of the medical condition observed by radiologists (Huang et al., 2021).

## 2. State of the art

### *2.1. Contrastive learning approaches*

Several recent works to utilize both medical images and text in the domain of chest X-ray (Huang et al., 2021; Li et al., 2021; Wang et al., 2022; You et al., 2023), using CLIP-based architecture. GLoRIA framework by Huang et al. (2021) uses an attention mechanism by contrasting image sub-regions and words in the paired report by learning attention weights that emphasize significant image sub-regions for a particular word to create context-aware local image representation. MedCLIP by Wang et al. (2022) on the other hand used unpaired images, text, and labels to enhance medical multi-modal learning. However this makes it less capable of retrieving the exact report for a given image due to the effect of decoupling image-text pairs, and as their approach relies on the performance of their rule-based labeler, it is not scalable to other diseases that the labeler can't address (You et al., 2023).

DeCLIP by Li et al. (2021) introduced a novel paradigm for data efficient CLIP that tackles the limitation of training data availability similar to the amount that CLIP was trained on through (1) self-supervision within each modality, (2) multi-view supervision across modalities, and (3) nearest-neighbor supervision from other similar pairs. CXR-CLIP by You et al. (2023) utilizes both image-text pairs not only from image-text dataset, but also from image-label dataset, thus tackles the lack of image-text data in the chest X-ray domain by expanding image-label pair via general prompting. In their work, they also used Multi-View Supervision (MVS) as inspired by Li et al. (2021), utilizing multiple images and texts in a chest X-ray study, such as two distinct images and texts pairs each using an augmentation approach.

### *2.2. Convolutional neural network approaches*

Other approaches have utilized convolutional neural networks in generating medical image descriptions or reports (Jing et al., 2017; Kisilev et al., 2016; Wang et al., 2018). In the work of Kisilev et al. (2016), they trained a CNN-based architecture to generate and rank rectangular region of interests of breast mammography and ultrasound modalities, where highest score candidates are fed to the subsequent network layers, in which they are trained to generate semantic description of the remaining ROI's. Their network is based on Faster R-CNN architecture (Ren et al., 2015), and was trained on mini-batches of positive and negative ROI candidates, and requires rectangular ground truth bounding boxes. Their main goal was to test the description stage of images using some descriptors such as mass shapes and margins.

Other approaches as Jing et al. (2017) utilized a hierarchical Long-Short Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997), apart of a multi-task learning framework to generate long report paragraph in chest X-ray domain. TieNet by Wang et al. (2018) is a multi-purpose text-image embedding network that utilizes report data together with paired images to produce meaningful attention-based image and text representations in the chest X-ray domain. Their approach also uses the paired text-image representations from training as a *priori* knowledge injected, to improve classification and generate text reports. They introduced an attention encoded text embedding mechanism that provides more meaningful text embedding, tackling the challenge that comes along with long reports of multiple information.

### *2.3. Limitations of current methods*

Despite such novel contributions made in the medical imaging chest X-ray domain using medical image-text datasets, several challenges still exist in the mammography X-ray domain, and specifically for BI-RADS report generation, which are summarized as follows:

- **Complications of mammography text reports**. Most of the present work utilizes chest X-ray image-text datasets, where the paired reports could be summarized under "impressions" and "findings", making it easy to extract text information for training. Mammography text reports on the other hand could contain additional information to the X-ray radiologist report that can be used as gold standard confirmation, such as ultrasound, MRI, or

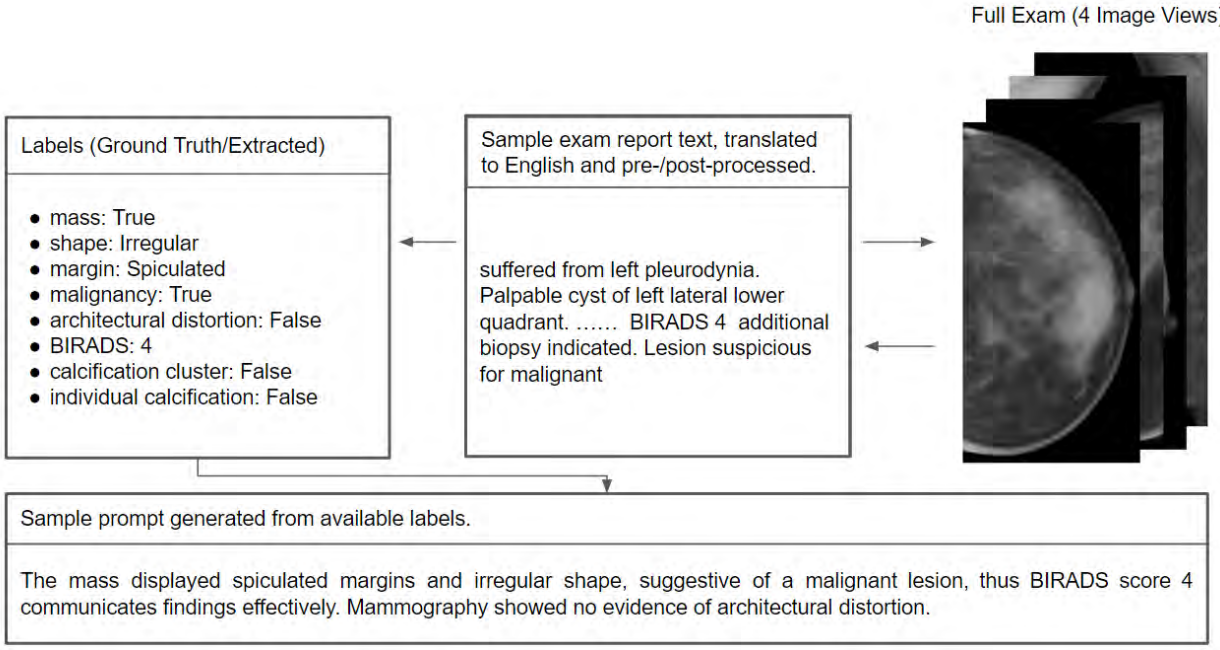



Figure 1: Example of mammography visual exam paired to different structure of text information, such as text report, extracted labels, or a prompt generated sentences using a text template based on the available labels for the exam.

pathology reports. Those additional information introduces challenges in identifying the best section for training a network. For instance, mammography X-ray radiology report could indicate suspicious morphology for a study, however, malignancy is confirmed by a biopsy and from an MRI exam. Such information can be mentioned in the same report of a single exam study, making it more challenging for the network to understand the meaning of different sections available in the patient reports.

- **Pathology variability in different views**. Unlike the work that is presented by CXR-CLIP (You et al., 2023), which utilizes up to two views with augmentation, mammography X-rays could contain up to four views (two for each breast - mediolateral oblique (MLO) and cranial caudal (CC)). With that, it could be possible to have a specific pathology in one breast and not in the other, increasing the necessity of having a network that is capable to process all four exam views and pair them to the text dataset.

- **Limited available data**. Most image-text datasets which are publicly accessible are available for different domains as chest X-ray (Bustos et al., 2020; Johnson et al., 2019), unlike mammography X-ray. And as the nature of its radiology reports, it is even more difficult to find paired images and full text reports, leaving a vast majority of image-label datasets unused to tackle the report generation task.

### *2.4. Contributions of this work*

The main contribution of this work is summarized as follows:

1. To our knowledge, this is the first work to utilize CLIP approach in mammography X-ray domain for mammography report generation. We tackle the lack of data by utilizing image-label and exam-reports paired datasets, as well as generating text prompts based on available labels to support the training. Our method, namely *MMG-CLIP*, does not depend on a ruler-based labeler, and doesn't require bounding boxes or small-patched images for training, and can be adapted to any image-label or exam-reports dataset.
2. We implemented a training approach that utilizes four views per exam, pairing them to the same text description, whether a label, a generated text prompt, or a report used during training or evaluation.
3. Performance of our model is validated on multiple downstream classification tasks, using zero-shot and supervised classification settings, as well as measuring the performance with respect to data-efficiency.
4. We introduced the report generation pipeline as a series of zero-shot classification tasks following BI-RADS guidelines, to obtain a clinical meaningful draft report for the patient exam.

## 3. Material and methods

The aim of this work is to learn a multi-modal embedding space from features that are extracted from an image and text encoders, and projected to a similar embedding dimension, to maximize the cosine similarity of both image and text embedding of real pairs in each batch, and minimize the cosine similarity of the incorrect embedding pairings, similarity to CLIP (Radford et al., 2021). Our approach aims to learn the image level or exam level characteristics of the 2D mammography X-ray images, up to four image views per exam. Those characteristics are also sampled from both image-label and exam-report datasets, in addition to the prompt generation approach to support training. In the following subsections, we further explain our work.

### *3.1. Data Sampling*

To train the model, each batch consists of both visual and textual information. Similarly to the work presented by You et al. (2023), we utilize a set of images, however, each exam could contain to up to four image views. Thus, each batch sample consists from one to four $X_{\text{img}}$ images depending on their availability for each exam, and $T_{\text{txt}}$ text. To simplify the following explanation, we denote quantities related to the full exam as $X_{\text{exam}}$ as in equation 1.

$$X_{\text{exam}} = \{X_{\text{img}}\}_{\text{img}=1}^{4} \tag{1}$$

In the case of image-label dataset, the sampled text $T_{\text{txt}}$ could be the an exact single label, for instance "benign" or "malignant" labels. Also, we use such labels, with any other labels found for the image to generate prompts that supports the model training. Those prompts we used contains more than one class label information, unlike the work of You et al. (2023) that only consists of one class-specific information. We also considered cases where the image-label pairs are missing labels information, making the prompts close to real clinical reports and taking into account not only the class information but their appearance.

For the exam-report dataset, the sampled text $T_{\text{txt}}$ consists of the processed report information, using certain selected reports sections found in the report text. In addition to that, as we had labels for the exams, we also experimented the training performance with generated sentences based on labelled data, known as prompts, and with both reports and prompts combined. We demonstrate a sample of image-prompt pairs from the training set in Appendix A, where we used our prompts as text input for training. We also took into account that those prompts are applicable with the BI-RADS guidelines and information that can be extracted from it. Figure 1 demonstrates different types of mammography datasets. Further details on the dataset and prompting mechanism is elaborated in subsection 4.1.

### *3.2. Model Architecture*

Motivated by CLIP by Radford et al. (2021), we proposed slight modification to how the embedding are extracted from multiple exam views to allow processing more than one mammography X-ray image at one time, as well as text feature extraction, both are described in subsections 3.2.1 and 3.2.2. In subsection 3.2.3, we describe the projection approach, that is necessary to align the embedding to the same dimension. Finally, subsection 3.3 describes the loss term that trains the model. All of this is summarized in Figure 2.

#### *3.2.1. Image Encoder*

The image encoder was used extract features from each exam input image, where the encoder can be referred to as in the following equation 2.

$$x = E_{\text{img}}(X_{\text{img}}) \tag{2}$$

where $x \in \mathbb{R}^{1\times D_{\text{img}}}$ represent the feature vectors for a single image view, and $E_{\text{img}}$ represents the image encoder. This is repeated for $N$ number of exam views, denoted as $x_{\text{exam}}$ where $x_{\text{exam}} \in \mathbb{R}^{N\times D_{\text{img}}}$. The value $D_{\text{img}}$ is the dimension of each vector. To obtain an overall visual representation of the exam, we average the values of all feature vectors of all exam views along the 0-th dimension, denoted at $x_{\text{f}}$, which is computed as following.

$$x_{\text{f}} = \frac{1}{N}\sum_{i=1}^{N} x_{\text{exam}}(i) \tag{3}$$

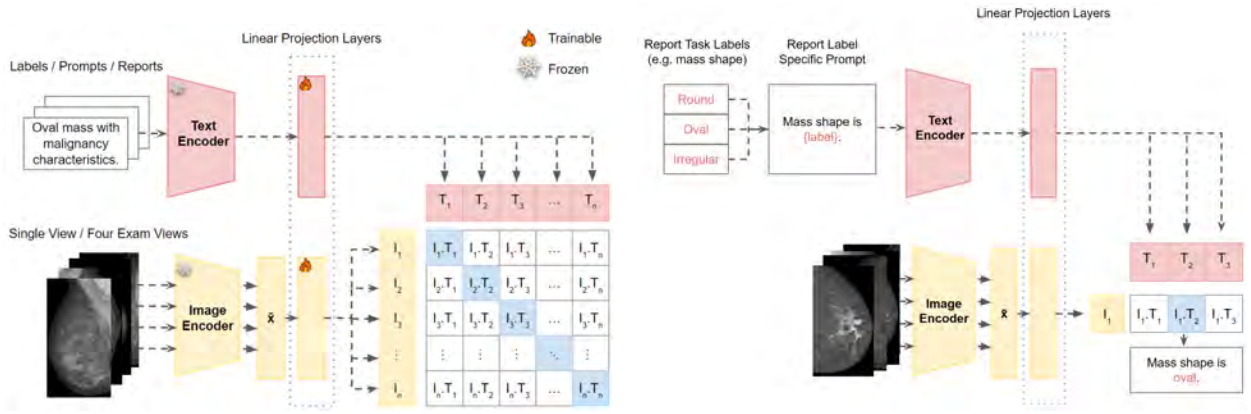
where $x_{\text{exam}}(i)$ represents the $i$-th column of matrix $x_{\text{exam}}$. The resulting $x_{\text{f}}$ has shape $(1, D_{\text{img}})$ representing the final image embedding vector. In the case that the network is trained at the image level where the input consists of a single image paired with the text, the averaging process is not performed and equation 2 is denoted as $x_{\text{f}}$.

The image encoder we used is a ConvNeXt Tiny model (Liu et al., 2022), pre-trained on an internal multi-vendor dataset from Fujifilm, GE HealthCare, HOLOGIC, Lorad, Philips and Siemens Healthineers, on large-scale dataset (>100K exams) for malignancy classification. In addition to that, we used ResNet-50 model from He et al. (2016) with ImageNet weights pre-trained on ImageNet tasks (Deng et al., 2009) in our ablation study to assess the performance when using a domain-specific pre-trained model to other pre-trained models.

#### *3.2.2. Text Encoder*

The text encoder was used to extract features from the input text. It can be described as the following equation 4.

$$t_{\text{f}} = E_{\text{txt}}(T_{\text{txt}}) \tag{4}$$




(a) Training Architecture

(b) Report task zero-shot prediction

Figure 2: Summary of our approach motivated by CLIP (Radford et al., 2021). *MMG-CLIP* extracts features from both text and image view/exam views, averages the image embedding, and projects them to predict the correct pairings of each batch. At inference, the network outputs unnormalized probability distribution for the input texts representing their probability to be paired to the input image. We aim to utilize this approach in report generation where a draft report is generated as a sequence of zero-shot classification tasks based on BI-RADS guidelines.

where $t_f \in \mathbb{R}^{1\times D_{txt}}$ represents the text embeddings and $E_{txt}$ represents the text encoder. We used BioClinicalBERT model by Alsentzer et al. (2019), which is a Bidirectional Encoder Representations from Transformers (BERT) based model as our text encoder, that was pre-trained using clinical dataset MIMIC-III (Johnson et al., 2016), similar to (Huang et al., 2021; Wang et al., 2022; You et al., 2023).

We also used BiomedBERT previously named as PubMedBERT (Gu et al., 2021), and BioGPT by Luo et al. (2022) to compared the performance when using BioClinicalBERT in our ablation study as in section 5. BiomedBert is also a variant of BERT models (Devlin et al., 2018), that was pre-trained from scratch on data collection from PubMed [1] that consists of 14 million abstracts and 3.2 billion words. This model was pre-trained on biomedical domain-specific data compared to BERT that is trained on Wikipidia [2] and BookCorpus (Zhu et al., 2015) as cited in (Gu et al., 2021). BioGPT is a variant of GPT large language models (LLMs), that is a domain-specific generative Transformer language model pre-trained on large-scale biomedical literature for biomedical text generation and text mining (Luo et al., 2022). It was pre-trained on 15M PubMed abstracts from scratch on GPT-2 (Radford et al., 2019) model configuration as a backbone, thus resulting into a model with 0.355 billion parameters in total as cited in Luo et al. (2022). In our experiments, we used all of the pre-trained text encoders from from HuggingFace [3].

#### 3.2.3. Embedding Projection

To align both the image embedding $x_f$ and text embedding $t_f$ in the same multimodal feature space, we trained linear layers as projection heads.

$$v = \frac{f_x(x_f)}{\|f_x(x_f)\|} \tag{5}$$

$$u = \frac{f_t(t_f)}{\|f_t(t_f)\|} \tag{6}$$

where $f_x$ is the projection head for the image embedding, $f_t$ is the projection head for the text embedding, $v$ and $u$ are the normalized projected embedding, $V = \{v\}_{i=1}^{n}$, $U = \{u\}_{i=1}^{n}$, and $n$ is the batch size.

### 3.3. Loss Function

For the loss, CLIP utilizes InfoNCE loss by Oord et al. (2018) as cited in Li et al. (2021), which is a symmetrical loss for image and text encoder. It iteratively trains both image and text encoders to maximize the cosine similarity of the image and text embedding of the $N$ real pairs in the batch, while minimizing the the cosine similarity of the image and text embedding of the $N^2 - N$ incorrect pairs (Radford et al., 2021). This is done by maximizing the alignment between both image-text pair, pulling their embedding closer, versus random pairs, pushing their embedding farther in the embedding space. This loss consist of maximizing the posterior probabilities of image embedding given its corresponding text embedding and the other way around , this way it ensures that the image-text correlation is asymmetric to either modality.

The loss for the image encoder can be denoted as in Equation 7, where as the loss for the text encoder can be denoted as in Equation 8.

$$L_I(U,V) = -\frac{1}{n}\sum_{u_i\in U}\log\left(\frac{\exp\left(\frac{v_i^T u_i}{\tau}\right)}{\sum_{v_j\in V}\exp\left(\frac{u_i^T v_j}{\tau}\right)}\right) \tag{7}$$

[1] https://pubmed.ncbi.nlm.nih.gov/
[2] https://www.wikipedia.org/
[3] https://huggingface.co/

$$L_{\mathrm{T}}(U,V) = -\frac{1}{n}\sum_{v_i \in V} \log\left(\frac{\exp\left(\frac{u_i^T v_i}{\tau}\right)}{\sum_{u_j \in U}\exp\left(\frac{v_i^T u_j}{\tau}\right)}\right) \tag{8}$$

where $\tau$ is a learnable temperature to scale logits, and it is fixed to 0.07. It controls the range of the logits and is directly optimized during training as a log-parameterized multiplicative scalar to avoid turning as a hyper-parameter (Radford et al., 2021). The similarity between the projected image embedding $v_i$ and text embedding $u_i$ is measured by the dot product between the embeddings.

The overall loss for a batch of image or exam and text pairs using $U, V$ notations can be described as the average of $L_{\mathrm{I}}$ and $L_{\mathrm{T}}$ as in Equation 9.

$$L_{\mathrm{CLIP}}(U,V) = \frac{1}{2}\left(L_{\mathrm{I}} + L_{\mathrm{T}}\right) \tag{9}$$

### *3.4. Interpreting Model Predictions and Outputs*

At prediction, our network outputs logits, which are unnormalized predictions, for each input text prompt as shown in Figure 2b. We normalized the logits to obtain normalized probabilities using a *softmax* layer, and thus we match the text prompt with the highest similarity as the correct prediction to the input image or exam. Figure 3 shows different evaluation examples we generated on different classification tasks using the same input image and different input text.

### *3.5. Evaluation Procedure*

We evaluated our implementation based on the experiments defined in Table 1, using both supervised classification and zero-shot classifications settings. The objective of comparing our image-label model trained on malignancy classification to the same encoder used in the network, which is a CNN, was to ensure that the model is able to perform an easy binary or multi-class classification task, thus we evaluated it using supervised approach. We reported the Binary Area Under ROC (AUROC) curve for binary tasks, and average AUROC with standard deviation for multi-class tasks.

We then added more complexity in terms of visual information or textual information (generated prompts sentences or reports or both combined) and measured the performance using zero-shot classifications using a class-specific generated prompts, as demonstrated in Figure 2b. We performed bootstrapping on 1000 samples, and averaged the AUROC of all of them, with the 95% confidence interval for binary tasks, and average AUROC with standard deviation for multi-class tasks. We also performed data-efficiency evaluation on different training data percentages for zero-shot evaluation. All experiments that uses single image as input will be referred to as "image level", whereas all experiments that uses an exam with several images will be referred to as "exam level".

We also demonstrated the benefit of utilizing projection layers on top of the encoders we used by plotting t-SNE by Van der Maaten and Hinton (2008) of the image embeddings.

### *3.6. Computational Resources*

All experiments were conducted on a NVIDIA TITAN V GPU with 12GB of memory. The code was implemented using PyTorch 1.13.1+cu116 in a Linux environment.

## 4. Experiments Results and Discussion

### *4.1. Datasets*

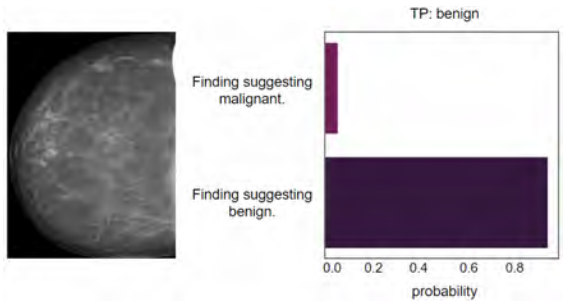
**Image-Label** dataset is annotated at the image level, consisting of one mammogram view and several annotation labels. At the high level, it consisted of 3311 benign annotated files, and 3174 annotated as soft tissue lesions (STL) files, making a total of 6485 samples. Those files contained other several region level annotations, such as architectural distortion, benign or malignancy, calcification cluster or mass, and properties such as histology, mass shape, mass margin, mass density, and subtlety.

Among all of the samples, we re-splitted the dataset into more image level labels, either benign or malignant. Those image views that were known as malignant, but has benign label were eliminated as they could be wrongly labelled. Thus a total of 3311 benign samples, and 1653 malignant samples, with their internal region level annotations. Table 2 summarises all of the labels we used from this annotated dataset. Any "unknown" label within this table means that the label was missing in the original dataset.

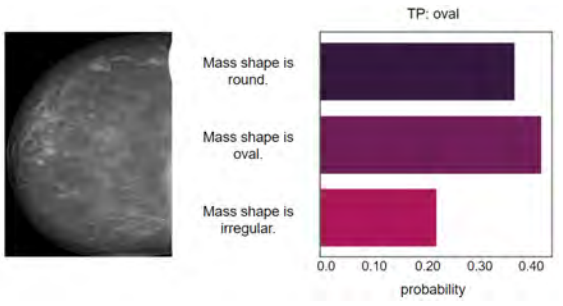
Another internal annotated dataset that was used consisted of 9696 ground truth annotations for other image views samples (or included). This dataset consisted of several annotations such as malignancy, asymmetry, calcification, mass, histology, biopsy and several others.

**Exam-Reports** is an internal dataset that contains four image views per exam (or less views if they were not collected or available), and a long Dutch report. It consists of 10,801 exam-report samples. Among all of those samples, only 1832 were applicable to be used, excluding several pathology, biopsy, or duplicates and only selecting mammogram reports. We also extracted labels from the sentences and manually translated them to their English labels found in BI-RADS guidelines to minimize the translation error.

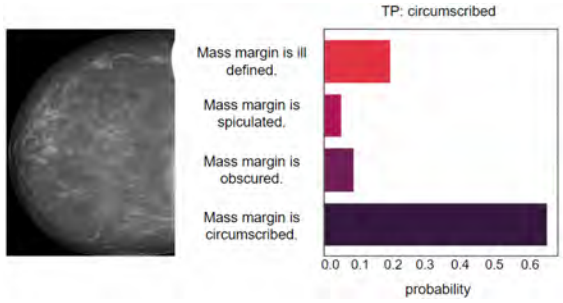
**Multi-label Prompts** are sentences generated randomly that contain one or more labels information. These sentences are formed by randomly selecting a template sentence describing each label, and concatenating them to form one or more sentences describing



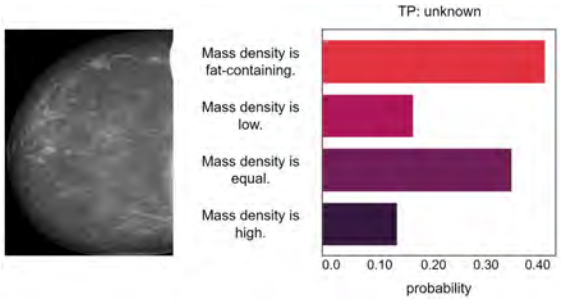

(a) Model output on two malignancy evaluation prompts




(b) Model output on three mass shape evaluation prompts




(c) Model output on four mass margin evaluation prompts




(d) Model output on four mass density evaluation prompts

Figure 3: Demonstration model inference output on four different examples, where all of the output similarities are normalized. We run different inference text prompts on the same input image. In the figures, *TP* stands for True Positive.

the image or exam. Thus, forming a structured paragraph used to train the network. The labels used for generating the prompts are from any of the labelled datasets, and the additional labels extracted from the reports. The prompts text can be paired to either image or exam level datasets, as explained in Table 1. The process of generating the prompts can be found in Appendix B.

Table 3 summarizes the split of the datasets used for training, validation, and testing, where it was (70%, 15%, and 15%) respectively. To make the results comparable, the exam-reports dataset test split was the exact same test split for the image-label datasets.

### *4.2. Baseline*

**ConvNext Tiny** model (Liu et al., 2022), that is the same model used as an image encoder in our approach. This encoder will be used as the baseline for malignancy detection, when comparing to our models trained on image-label experiment dataset.

### *4.3. Implementation Details*

For the visual information, both at image and exam levels, we did not perform any augmentation or preprocessing. As text reports were originally in Dutch language, we translated them after pre-processing to standardise the training in English using the command =GOOGLETRANSLATE(text_column, "nl", "en") in Google Spreed Sheets [4]. Pre-processing included eliminating unnecessary reports samples, text cleanup that includes cleaning redundant words, structures, spaces, special characters, or patterns. As the nature of the mammography reports could include additional gold standard information that assist in evaluation of abnormalities, such as current study, ultrasound, mammogram X-ray, MRI, pathology, we selected only three types that we found contains most of the important information, that are current study, mammogram X-ray, and MRI. This was also performed during the preprocessing. The post-processing of the text was performed after the translation mainly to remove any duplicate sentences within the text, as the performance will heavily rely on the translation performance.

As for the embeddings, the final image and text embedding sizes are 512. Both encoders were frozen and only linear layers were trained on top of them. For both image-label (either binary or multi-class) and image-prompts experiments training, we used 1 linear layers with a ReLU activation function and dropout layer. By experimenting, we used dropout of 0.2 and 0.5 for the image-label and image-prompts training respectively. For any of the exam level experiments, we used a 2 trainable linear layers. For the training, we tracked the validation loss curves and several other area under the ROC (AUROC) values.

For all of the experiments, the early stopping condition was set with patience of 5 monitoring the validation loss and a tokenizer sequence length of 256. For the hyper-parameters, we used a cosine-annealing learning-rate scheduler (Loshchilov and Hutter, 2016), with a warm-up epoch of 0.1 and 30 trainable epochs, AdamW (Loshchilov and Hutter, 2017) optimizer with an ini-

[4] `https://www.google.com/sheets/about/`

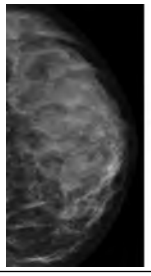
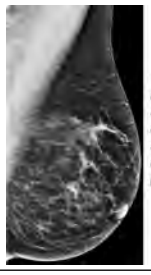
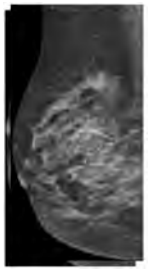
| Experiment Name | Description | Input Example |
|---|---|---|
| Image-Label | Training with images and labels. |  "benign" |
| Image-Prompts | Training with images and prompts generated. |  "Imaging revealed a mass with spiculated margins and irregular shape, suggestive of malignant pathology." |
| Exam-Reports | Training with exams and reports text. |  "Status after amputation of left breast due to carcinoma. Palpable abnormality on the right at 10 o'clock of 1.5 cm with skin retraction….." |
| Exam-Reports + Prompts | Training with exams and reports text combined with prompts. |  "The mass was characterized by ill defined margins and oval shape on imaging, suggesting a potential malignant etiology, assigning BIRADS score of 4 based on the findings. Fast growing tumor of right breast/axilla. Excision in February followed by recurrence (PA of excision inconclusive). Currently malignant …." |
| Exam-Prompts | Training with exams and prompts generated. |  "The mass displayed spiculated margins, suggestive of a malignant lesion, the mammography report assigns a BIRADS score of 5 to guide further clinical decisions." |

Table 1: Experiments description and the datasets used in each of them.

tial learning rate 5e-5, **[EOS]** token's final output as the global textual representation, and weight decay 1e-4 following the work of You et al. (2023). For image-label experiments, we used a batch size of 32 samples for all three splits, whereas for the remaining experiments, we used batch size of 64.

### *4.4. Classification*

We started by evaluating the learned representation on several image classification tasks based on our image-label dataset available labels mentioned in Table 2, using both supervised image classification and zero-shot classification settings. In both settings, as mentioned earlier, we only trained linear projection layers on top of the pre-trained encoders.

#### *4.4.1. Supervised Image Classification*

For the supervised classification, as our baseline CNN encoder was pre-trained on malignancy task, we trained our network on the malignancy labels of the

| Label Group | Labels Names | Count |
|---|---|---|
| Malignancy | Benign | 3311 |
| | Malignant | 1653 |
| Mass Margins | Unknown | 2467 |
| | Ill defined | 1095 |
| | Obscured | 697 |
| | Spiculated | 484 |
| | Circumscribed | 221 |
| Mass Shapes | Unknown | 2466 |
| | Irregular | 1218 |
| | Round | 681 |
| | Oval | 599 |
| Architectural Distortion | Normal | 4842 |
| | Distortion | 122 |
| Calcification | No Calcification | 2969 |
| | Has Calcification | 1995 |
| Mass | No Mass | 278 |
| | Mass | 4686 |

Table 2: Image-Label dataset description.

| Dataset | Split | Count |
|---|---|---|
| Image-Label *or* | Train | 3474 |
| Image-Prompts | Valid | 1490 |
| | Test | 745 |
| Exam-Reports *or* | Train | 1282 |
| Exam-Reports + Prompts *or* | Valid | 550 |
| Exam-Prompts | Test | 745 |

Table 3: Datasets split summary. First row summarizes the image level splits, either using labels or prompts depending on the experiment, and second row summarizes the exam level splits.

image-label dataset, and compared the results area under the ROC curve (AUROC) of the true class. We also trained a network for the other labels of the dataset and reported the results in Table 4. In our results, we show that our network was able to outperform a traditional CNN performance on malignancy detection by training a single linear layer. Our network also performed well on the remaining classification tasks. The main objective was to ensure that the network is capable of learning a simple label classification task, either binary or multi-class using the learned representation from both image and text modalities.

#### *4.4.2. Zero-shot classification*

For the zero-shot prompt classification, the network was trained and evaluated on different experiments, thus different representations. The constructed evaluation text prompts were specified to target the model performance in understanding the clinical meaning of the text input as a full sentence. Therefore, we constructed a class-wise inference prompt for each label task. Those inference prompts are different from the prompts generated for training, and can be found in Table 6. We evaluate the binary classification tasks by computing the AUROC of 1000 bootstrapped samples with 95% CI, and computed the average AUROC for multi-class classification tasks with standard deviation. We also evaluated the performance on both datasets, at image and exam level training, and to make the evaluation fair, all experiments were evaluated on the same test samples at the image level.

As shown in Table 5, both experiments image-prompts and exam-prompts outperform all other experiments, where those experiments were trained on different dataset samples, and on the same text prompting approach we proposed. Training the network with well structured sentences as the generated prompts performs better than training with real radiologist reports as the nature of the text reports when they are written, they are not generally standardised. This can be also demonstrated when training the network with exam-reports and exam-reports + prompts, where including the prompts improved the results as demonstrated in the table. It is also worth noting that each experiment row in Table 5 is a single model performance, thus shows the ability in generalizing to different downstream tasks.

### *4.5. Data-efficiency Evaluation*

We further evaluated the model performance for zero-shot classification taking into account different sizes of training dataset samples (10%, 20%, 50%, and 100%), on malignancy detection. In Figure 4, we show that both of our models, either trained on image-label malignancy task, or on exam-prompts experiments improve the performance when more training data is used, tracking their malignancy AUROC metric for all of the test samples. The image-label trained model shows only slight improvement as the encoder only performance (in red color) is high, so training linear layers on top of the pre-trained encoder improves its ability in malignancy zero-shot classification for this specific dataset. It demonstrated a consistent high performance on all percentages of the training data. The exam-prompts model that is trained on more visual and textual information showed a significant improvement in the malignancy zero-shot detection with different percentages, indicating that the model is effectively learning from the additional data.

### *4.6. Report Generation*

To generate a radiology report, we defined a report as a series of zero-shot classification tasks. Those can be specific based on BI-RADS mammography guidelines, or general to any other inference task. To generate a report, we used the exam-prompts experiment model, and constructed a series of inference tasks. The final step of the report generation includes formatting all outputs into a template sentences and concatenating the results to form a single report. In Figure 5, we demonstrate a summary of our report generation pipeline. At the top

| | Binary AUROC ↑ | | | | Average Multi AUROC (± std) ↑ | |
|---|---|---|---|---|---|---|
| Experiments | Malignancy | Arch. Dist. | Mass | Calcification | Mass Shapes | Mass Margins |
| CNN (Baseline) | 0.9153 | - | - | - | - | - |
| Image-Label | 0.9402 | 0.8293 | 0.8005 | 0.8820 | 0.8023 (± 0.078) | 0.8344 (± 0.089) |

Table 4: Comparison of area under the ROC (AUROC) of different experiments and classification tasks (binary and multi-class) using one-vs-all classification evaluation on image level experiments. Total 745 samples of the image-label dataset test split were used. In the table headers, Arch. Dist. stands for architectural distortion.

| | Average Binary Bootstrap Samples AUROC (95% CI) ↑ | | | | Average Multi AUROC (± std) ↑ | |
|---|---|---|---|---|---|---|
| Experiments | Malignancy | Arch. Dist. | Mass | Calcification | Mass Shapes | Mass Margins |
| Image-Prompts | **0.931 (0.905-0.953)** | 0.682 (0.554-0.808) | 0.663 (0.564-0.755) | 0.680 (0.639-0.719) | 0.727 (± 0.120) | **0.715 (± 0.154)** |
| Exam-Reports | 0.828 (0.791-0.861) | 0.637 (0.504-0.78) | 0.475 (0.3721-0.572) | 0.567 (0.524-0.610) | 0.596 (± 0.079) | 0.560 (± 0.089) |
| Exam-Reports + Prompts | 0.847 (0.814-0.878) | 0.646 (0.509-0.791) | 0.527 (0.425-0.619) | 0.683 (0.644-0.723) | **0.848 (± 0.088)** | 0.594 (± 0.094) |
| Exam-Prompts | 0.916 (0.891-0.938) | **0.717 (0.620-0.804)** | **0.678 (0.603-0.743)** | **0.736 (0.701-0.772)** | 0.700 (± 0.106) | 0.639 (± 0.218) |

Table 5: Comparison of the average area under the ROC (AUROC) of different experiments and classification tasks (binary and mutli-class) using zero-shot classification evaluation on both image and exam level experiments. For binary tasks, we bootstrapped 1000 samples, and computed the average AUROC and 95% CI. For the multi-class tasks, we computed the average AUROC ± standard deviation. Total 745 samples of the image-label dataset test split were used. In the table headers, *Arch. Dist.* stands for architectural distortion.

| **Label Group** | **Input Evaluation Prompt** |
|---|---|
| Malignancy | Findings suggesting {*label*}. |
| Mass Margins | Mass margins is {*label*}. |
| Mass Shapes | Mass shape is {*label*}. |
| Architectural Distortion | Normal architecture is visible.<br>Displayed architectural distortion. |
| Calcification | No calcifications are present.<br>Finding suggesting calcifications. |
| Mass | No mass was observed.<br>Findings revealed a mass. |

Table 6: Zero-shot evaluation prompts for all label groups. The {*label*} are replaced with the labels reported in Table 2, that are based on BI-RADS guidelines.

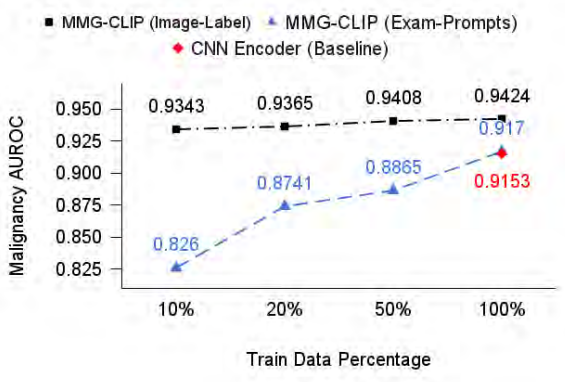



Figure 4: Image-label and exam-reports models (ours) zero-shot performance for malignancy classification using different amount of data, without bootstrapping.

level, an inference task is made to validate if an image or exam either has a mass, calcification, or no findings. As "No Findings" ends the report, it doesn't require any further evaluation for mass or calcification information, thus we report directly a conclusion sentence as shown in Figure 6b.

Both "Mass" and "Calcification" in Figure 5 have their own generation pipeline. In "Mass" track, we evaluate the malignancy, mass shape, mass margins, BI-RADS score, and architectural distortion. As for "Calcification" track, we evaluate malignancy, distribution, BI-RADS score, and architectural distortion. An example for a report generated for an exam with malignant mass findings is shown in Figure 6a, where as an example for a report generated for an exam with benign calcification is shown if Figure 6c.

One important limitation of our report generation is the decision condition taken for all prompts output similarities generated from the model. If the model fails on identifying the correct type of findings at the very first level of the generation pipeline, all following evaluation results will be wrong. Figure 6d shows a failed example of a report generated as "No Findings", where it contains other types of findings. As we take the maximum

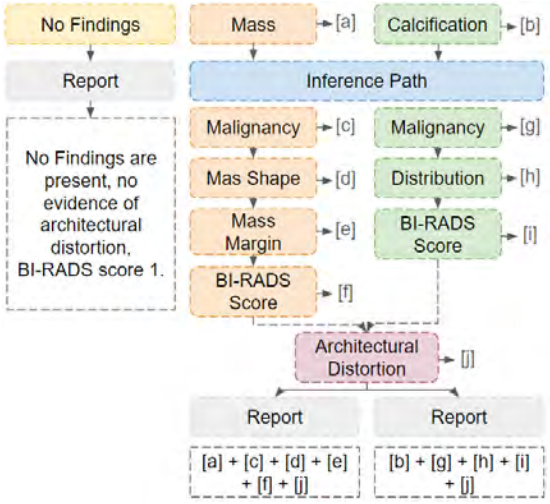



Figure 5: Report generation pipeline. Symbol *[letter]* represent the inference task output, and + represent output formatting and concatenation.

similarity value of all text-prompts output similarities, we are not able to distinguish between a strong prediction (with high probability for a specific text prompt) or for a confused prediction (when all probabilities are close to each others). Another concern is whether an exam or an image has more than one finding similar to both "Mass" and "Calcification" together. When taking the maximum similarity, we result with having only one output text to the inference task, thus can't combine multiple texts as an output.

### *4.7. Embedding Visualization*

Data visualization using dimension reduction approaches can assist in understanding the geometric and neighborhood structures of datasets (Wang et al., 2021). A popular tool to perform dimensional reduction is the t-distributed Stochastic Neighborhood Embedding (t-SNE) algorithm (Shah and Silwal, 2019), introduced by Van der Maaten and Hinton (2008), or principal component analysis (PCA). We performed t-SNE analysis on both the embeddings from the CNN encoder as in Figure 7a and 7c, as well as on the linear layers on top of the encoder as in our model in Figures 7b and 7d. The figures demonstrates the separation of "benign" and "malignant" embeddings classes of the malignancy classification task as an output of the networks projected into lower dimensional using t-SNE.

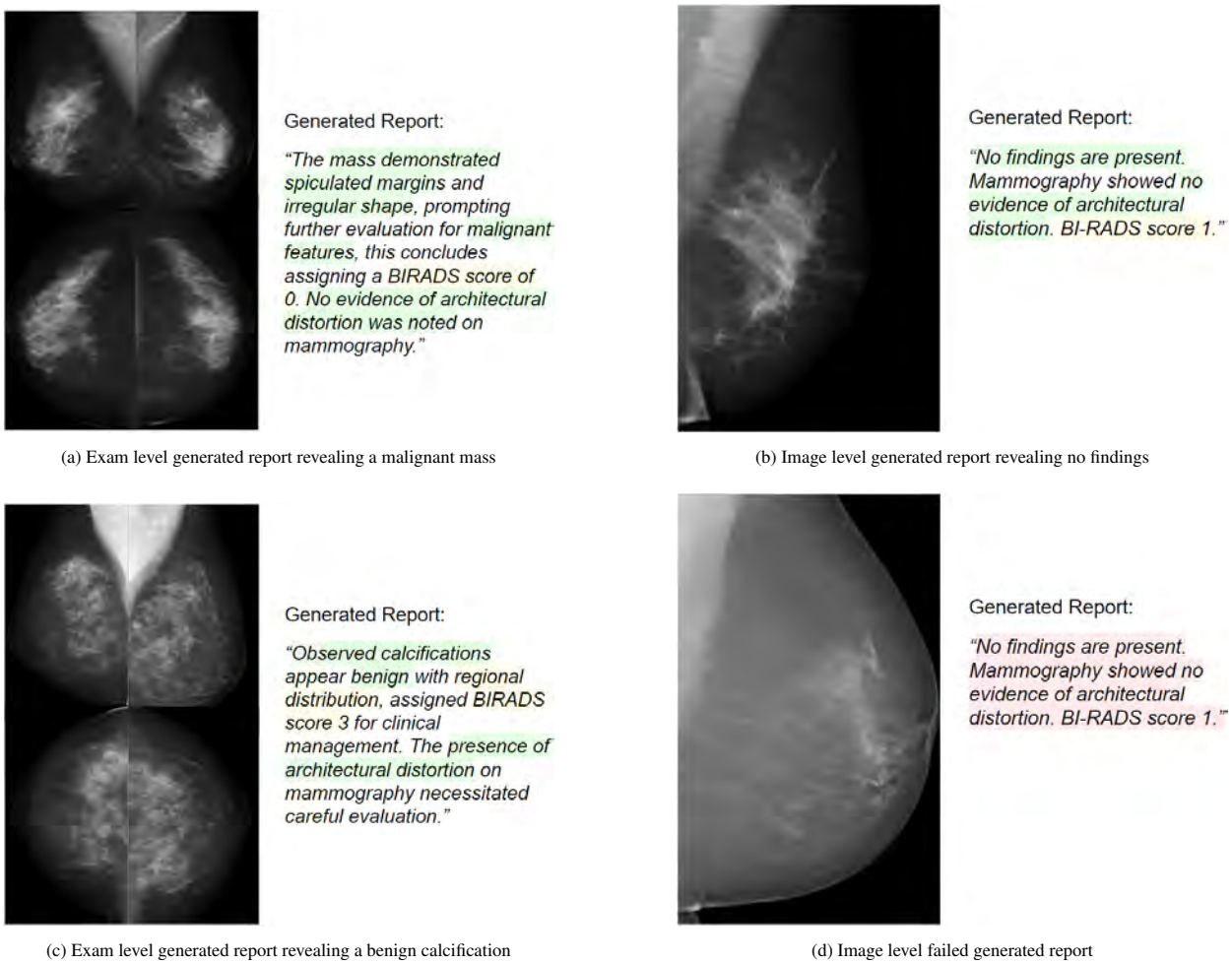
As shown in Figures 7a and 7b, both networks generates a well clustered points of both labels. Using the CNN only however shows some overlap between the two classes, indicating that the baseline CNN encoder does not completely distinguish between them. Adding linear layers on top of the pre-trained encoder does slightly produce better clusters as it focuses on the specific characteristics and patterns present in our datasets, thus making it perform better on our test cases and provides more distinct clusters with less overlap. We also visualized the first dimension of t-SNE with respect to the models probabilities to belong to malignancy class as shown in Figures 7c and 7d. Both Figures indicates positive correlation between the t-SNE dimension 1 and malignancy probabilities, where the model with projection layers as in Figure 7d shows more distinct and reliable probability estimates for malignancy, as there is a clearer separation between both labels cases compared to the baseline encoder alone in 7c.

### *4.8. Limitations and Future Work*

Despite that we reached promising results in our experiments, we believe that there are improvements that can be made.

**Embedding pairing** is a challenging task in medical image-text datasets as the nature of the visual and textual information can be paired to more than one sample. For example, an image can contain several regions of interest, where it can be described correctly in two separate reports sentences of two different exams. This makes the loss metrics not meaningful when it comes to training as pairing a single image-text pairs might not be meaningful when the network learns the global representation of all of the input data. This also was observed when training with large batches (that are possible to have reports with similar information) on a small datasets like ours, but not observed when using a very small batch size as the possibility of having two samples of same findings is much lower. We experimented implementing different variations of CLIP InfoNCE loss taking into account the batch samples and other sampling mechanisms to tackle the problem, however none of the approaches we tried proved better learning when when it comes to long text reports. Thus, a meaningful contrastive loss would be very beneficial for the network to be able to match medical image-text datasets. For example, region-wise matching between image and text information, or giving more weights to certain regions could potentially improve the network loss mechanism.

**Report generation**. When it comes to generating report, as we mentioned earlier we use the maximum similarity output as the final task result before creating a report. Trying different decision making approaches could be useful in generating more precise reports, but it also requires human intelligence and clinical validation. One case that we noticed could be failing repetitively is when network received 5 input prompts, and the five similarities values are very close to each other, using the maximum value might not be ideal. Also, some report details have more importance than others, for example malignancy classification, or differentiating between the presence of mass, calcification, or no findings, compared to other sub-tasks like mass region or calcification distribution. The decision making here plays an important role in the report accuracy, and taking the




(a) Exam level generated report revealing a malignant mass

(b) Image level generated report revealing no findings

(c) Exam level generated report revealing a benign calcification

(d) Image level failed generated report

Figure 6: Demonstration of report generation using a full exam as in (a), and (c), or a single image view as in (b) and (d). Text highlighted in green is a correct prediction from the network, where text highlighted in red are wrong predictions. Yellow highlighted text has no label to compare with.

maximum similarity value might not always be the best case. Thus, other decision making approaches such as applying a threshold value to the similarities could be explored in future work.

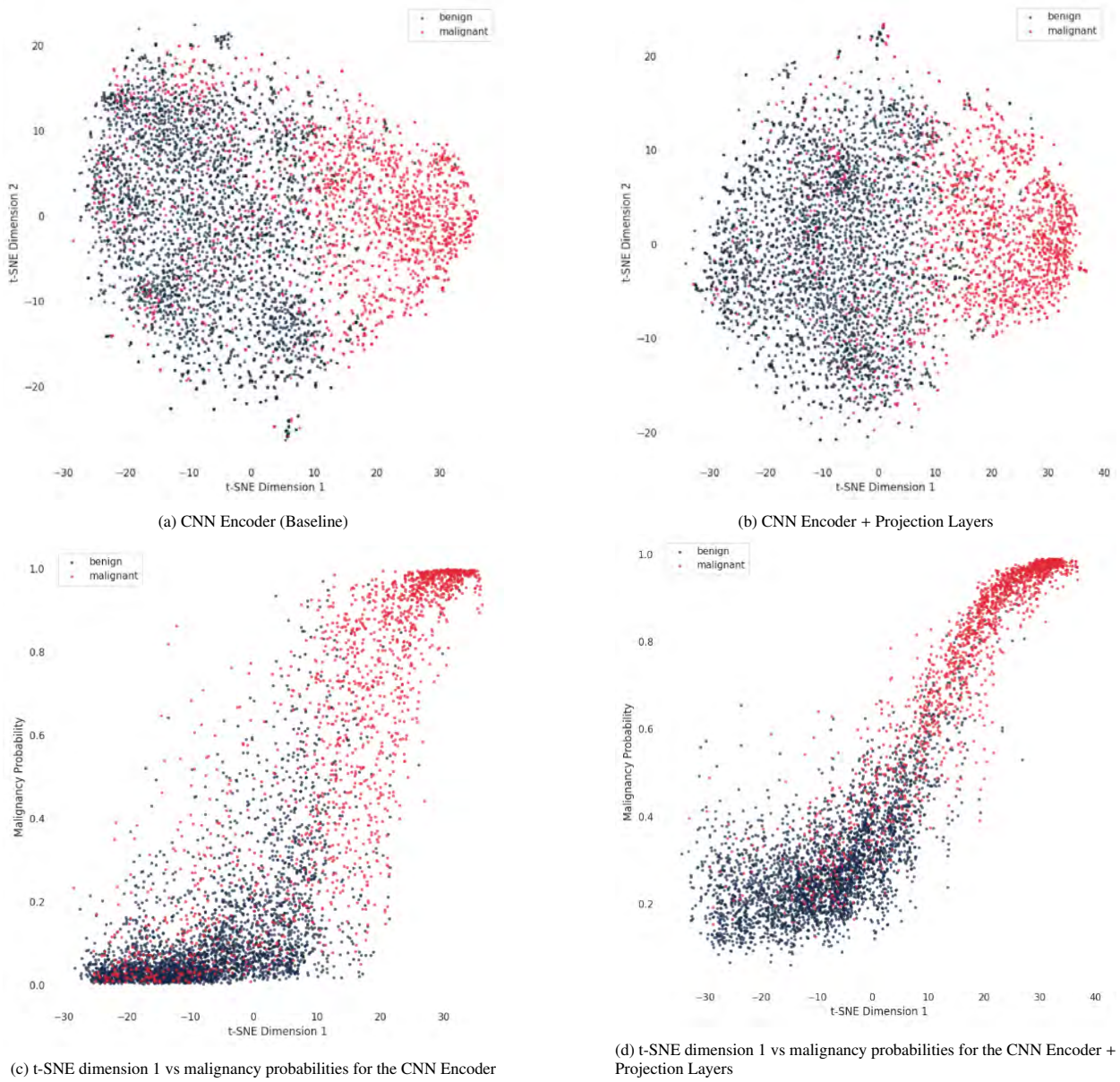
**Pre-training** the encoders on large scale datasets could significantly improve the performance when and generalization of the model. As we used pre-trained encoders, the extracted features relies on their performance as well as on the performance of the trained linear layers. And as we had a very small amount of data to work on, we were not able to train the models from scratch.

**Network Architecture** can be improved to localize the presence of the pathology reported in the text to which exam view it is found in. This can significantly improve the reporting precision if the network is capable of identifying which view exactly has more importance. In our implementation, while training the network, we averaged the features extracted from each of the input image views, thus losing the anatomical location of the pathology it contains. For example, when a mass appears in the "right MLO image view" in an exam, we lose such information while averaging the embedding. Having that considered can also improve the feature extraction approach to assign more weights to important views and less to others.

## 5. Ablation Study

**Ablation on model architecture.** To understand the effectiveness of the architectural parameters and key components, we conducted ablation study using different parameters and components with respect to malignancy zero-shot classification performance. All results reported in Table 7 were trained using the best exam-prompts experiment model. We used different training configurations to evaluate their impact on zero-shot classification performance on one task.

In the first row, we evaluate different number of projection layers. From the reported results, 2 Linear Projection layers gave the best zero-shot performance for our model and no indication of increased performance when more trainable layers are used.




(a) CNN Encoder (Baseline)

(b) CNN Encoder + Projection Layers

(c) t-SNE dimension 1 vs malignancy probabilities for the CNN Encoder

(d) t-SNE dimension 1 vs malignancy probabilities for the CNN Encoder + Projection Layers

Figure 7: Image embedding visualization of malignancy Image-Label dataset for both CNN encoder (Baseline) alone ours that includes projection layers on-top of the encoders.

In the second row of Table 7, we used the default 2 projection layers with different training batch sizes. The default value we used was batch size *n=64* with tokenizer sequence length of 256 for exam-prompts experiment model where it obtained 0.916 (0.891-0.938). Both *n=32* and *n=128* showed no significant improvement on the performance as reported in the table. Similarly to the tokenizer sequence length in the third row, both sequence lengths 384 and 512 didn't improve the performance of our default value. In addition to that, we observed that using a logit scale $\tau = 0.07$ performs better than without performing scaling to the logits during training as in the last row reported in the table.

**Ablation on inference prompts.** As mentioned previously, our evaluation prompts contribute significantly to the results we obtained, as we believe it targets the clinical meaning behind the label we are evaluating. To measure the impact of changing the inference prompts during zero-shot settings, we experimented using CXR-CLIP by You et al. (2023) evaluation prompts for zero-shot and compared the results to ours. In this evaluation, we are not comparing our results to theirs, as it is using totally different datasets in different domains, but only comparing our model behaviour to different evaluation prompts. In CXR-CLIP, they used the prompts "{*classname*}" versus "*No* {*classname*}" for all labels they evaluate, for example "No oval" versus "oval" for "Mass Shapes" task, and then using prediction of the "{*classname*}" to generate the results. We noticed that this introduces a challenge for our network when

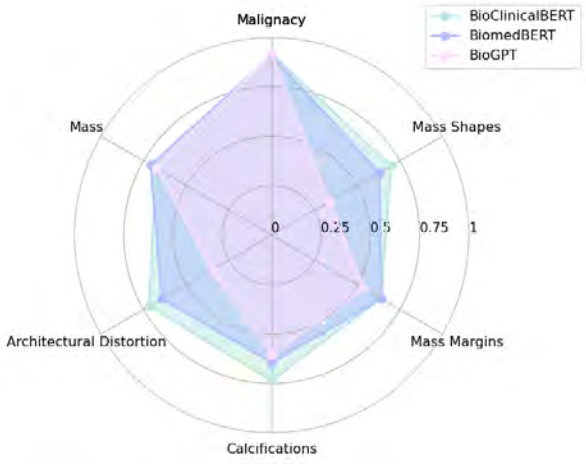
| Experiments | AUROC (95% CI) ↑ |
|---|---|
| MMG-CLIP | |
| w/ 1 proj. layers | 0.893 (0.864-0.920) |
| **w/ 2 proj. layers** | **0.916 (0.891-0.938)**[a] |
| w/ 3 proj. layers | 0.910 (0.882-0.933) |
| MMG-CLIP | |
| w/ batch size = 32 | 0.908 (0.883-0.933) |
| w/ batch size = 128 | 0.912 (0.885-0.936) |
| MMG-CLIP | |
| w/ seq. length = 384 | 0.910 (0.885-0.933) |
| w/ seq. length = 512 | 0.906 (0.877-0.929) |
| MMG-CLIP | |
| w/ logit scale = 1 (no scale) | 0.8876 (0.858-0.913) |

[a] Value obtained using the default experiment parameters as 2 proj. layers, batch size = 64, seq. length = 256, logit scale $\tau = 0.07$.

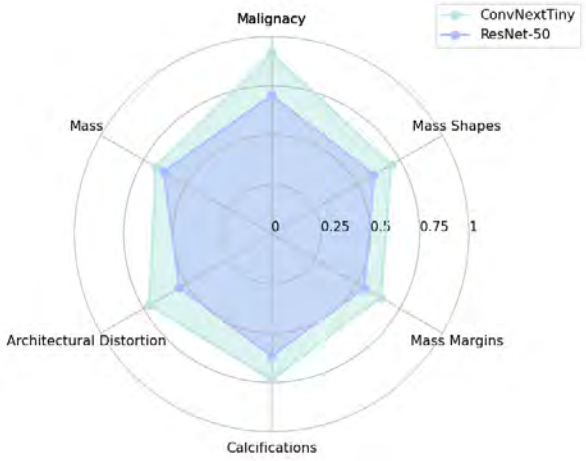
Table 7: Ablation study of key architectural parameters with respect to different parameters and components. The reported scores are the average AUROC of 1000 bootstrapped samples with 95% CI on malignancy zero-shot classification. In the table, *proj. layers* is projection layers, *seq. length* is the tokenizer sequence length.

| Experiments | AUROC (± std) ↑ |
|---|---|
| MMG-CLIP | |
| w/ CXR-CLIP prompts | 0.587 (± 0.074) |
| **w/ our prompts** | **0.700 (± 0.106)** |

Table 8: Ablation study of different evaluation prompts used to evaluate zero-shot settings. The reported scores are the average AUROC (± std) for all labels curves on "Mass Shapes" task.




(a) Different LLM models performance as text encoders.




(b) Different vision models performance as image encoders.

Figure 8: Comparison between using different vision and large language models as encoders in our network on zero-shot classification tasks on the exam-prompts experiment model. Values on the axis (0, 0.25, 0.75, and 1) are average AUROC values for 1000 bootstrapped samples for binary tasks, and average AUROC for multi-class tasks.

it comes to multi-class evaluation, where it performs poorly using their prompting mechanism compared to ours, for example "Mass shape is oval". Table 8 shows that with our evaluation, we obtain higher score for "Mass Shapes" task when using our prompts compared to using CXR-CLIP evaluation prompts.

**Ablation on pre-trained clinical text encoders.** As we used pre-trained text encoder BioClinicalBERT model by Alsentzer et al. (2019) and not pre-training our own due to the limited number of training data, the network performance heavily relies on the performance of the pre-trained text encoder. To understand the impact, we analyzed our network performance using other large language models of different parameter sizes as our text encoder. In Figure 8a, we compared the performance of our model trained using exam-prompts experiment, similar to the evaluation approach reported in Table 5.

As shown in Figure 8a, BioClinicalBERT model as our text encoder outperforms both BiomedBERT and BioGPT in performance for all zero-shot classification tasks. This supports idea of having a domain specific pre-trained model on clinical text datasets when it comes to learning medical text reports from other domains, and encourages pre-training a mammography specific text encoder for future work. Following BioClinicalBERT is BiomedBERT, where it shows a balanced performance across most tasks with particular strength for both "Mass" and "Malignancy". It also supports the idea that BERT variant models tends to outperform GPT variant models which are more commonly used in generation tasks (Luo et al., 2022). BioGPT was the least in performance as it had very low metric values, close to randomness. Thus, both BioClinicalBERT and BiomedBERT are more suitable text encoders encoding medical text given their performance on our various tasks, and could potentially be used in pre-training a BERT model for mammography domain-specific data.

**Ablation on pre-trained vision image encoders.** To assess the performance of a domain-specific pre-trained model as an image encoder such as our pre-trained ConvNeXt Tiny image encoder, we used a ResNet-50 model

and applied transfer learning approach to its last layer (layer 4), with similar training configurations. In Figure 8b, we show that having a pre-trained model on domain-specific knowledge significantly outperforms a model pre-trained on general vision task, where our ConvNext Tiny model performed better in all tasks. ResNet-50 model had consistent and balanced performance and was not biased to a specific task.

## 6. Conclusions

In this work, we proposed an image-text contrastive learning framework named *MMG-CLIP* as well as a report generation BI-RADS specific pipeline for mammography X-ray 2D images. Our implementation includes not only training the network at the image or exam level (multiple images) with medical text, but also utilises multi-class generated prompt text to improve the model performance on zero-shot classification tasks. *MMG-CLIP* showcases remarkable flexibility due the multi-modality and zero-shot learning ability. Our experiments results shows the network data-efficiency and zero-shot capability of the learned representations for various downstream classification tasks.

## Acknowledgments

I would like to express my deepest gratitude ScreenPoint Medical for providing me with this incredible opportunity to work on this thesis topic. First and foremost, I would like to thank my supervisors, Santiago Pires and Jaap Kroes for their guidance, support and mentoring. I'm also very grateful to all ScreenPoint teams for their assistant, feedbacks, and providing the necessary resources. Together, we achieved more than I could have ever accomplished alone. I would also like to extend my sincere appreciation to the MAIA master consortium for providing me with a solid educational foundation and to the European Commission for granting me this opportunity and generously funding my master education. Lastly, I would like to thank my family and friends for their endless support and unconditional love, despite the long distance.

## References

Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M., 2019. Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323 .

Anwar, S.M., Majid, M., Qayyum, A., Awais, M., Alnowami, M., Khan, M.K., 2018. Medical image analysis using convolutional neural networks: a review. Journal of medical systems 42, 1–13.

Bustos, A., Pertusa, A., Salinas, J.M., De La Iglesia-Vaya, M., 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis 66, 101797.

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., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

Fishman, M.D., Rehani, M.M., 2021. Monochromatic x-rays: The future of breast imaging. European Journal of Radiology 144, 109961.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H., 2021. Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) 3, 1–23.

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., et al., 2021. Pre-trained models: Past, present and future. AI Open 2, 225–250.

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.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.

Huang, S.C., Shen, L., Lungren, M.P., Yeung, S., 2021. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3942–3951.

Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T., 2021. Scaling up visual and vision-language representation learning with noisy text supervision, in: International conference on machine learning, PMLR. pp. 4904–4916.

Jing, B., Xie, P., Xing, E., 2017. On the automatic generation of medical imaging reports. arXiv preprint arXiv:1711.08195 .

Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S., 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data 6, 317.

Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G., 2016. Mimic-iii, a freely accessible critical care database. Scientific data 3, 1–9.

Kallenberg, M., Petersen, K., Nielsen, M., Ng, A.Y., Diao, P., Igel, C., Vachon, C.M., Holland, K., Winkel, R.R., Karssemeijer, N., et al., 2016. Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. IEEE transactions on medical imaging 35, 1322–1331.

Karimi, D., Dou, H., Warfield, S.K., Gholipour, A., 2020. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Medical image analysis 65, 101759.

Kisilev, P., Sason, E., Barkan, E., Hashoul, S., 2016. Medical image description using multi-task-loss cnn, 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. 121–129.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al., 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International journal of computer vision 128, 1956–1981.

Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., Yan, J., 2021. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. arXiv preprint arXiv:2110.05208 .

Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11976–11986.

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

Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 .

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., Liu, T.Y., 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. Briefings in bioinformatics 23, bbac409.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-sne. Journal of machine learning research 9.

Mohamed, A.A., Berg, W.A., Peng, H., Luo, Y., Jankowitz, R.C., Wu, S., 2018. A deep learning method for classifying mammographic breast density categories. Medical physics 45, 314–321.

Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 .

Pesapane, F., Trentin, C., Ferrari, F., Signorelli, G., Tantrige, P., Montesano, M., Cicala, C., Virgoli, R., D'Acquisto, S., Nicosia, L., et al., 2023. Deep learning performance for detection and classification of microcalcifications on mammography. European Radiology Experimental 7, 69.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR. pp. 8748–8763.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al., 2019. Language models are unsupervised multitask learners. OpenAI blog 1, 9.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 28.

Ribli, D., Horváth, A., Unger, Z., Pollner, P., Csabai, I., 2018. Detecting and classifying lesions in mammograms with deep learning. Scientific reports 8, 4165.

Salama, W.M., Aly, M.H., 2021. Deep learning in mammography images segmentation and classification: Automated cnn approach. Alexandria Engineering Journal 60, 4701–4709.

Shah, R., Silwal, S., 2019. Using dimensionality reduction to optimize t-sne. arXiv preprint arXiv:1912.01098 .

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Wang, J., Yang, X., Cai, H., Tan, W., Jin, C., Li, L., 2016. Discrimination of breast cancer with microcalcifications on mammography by deep learning. Scientific reports 6, 27327.

Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M., 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 9049–9058.

Wang, Y., Huang, H., Rudin, C., Shaposhnik, Y., 2021. Understanding how dimension reduction tools work: an empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. Journal of Machine Learning Research 22, 1–73.

Wang, Z., Wu, Z., Agarwal, D., Sun, J., 2022. Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 .

Xie, X., Niu, J., Liu, X., Chen, Z., Tang, S., Yu, S., 2021. A survey on incorporating domain knowledge into deep learning for medical image analysis. Medical Image Analysis 69, 101985.

Yala, A., Lehman, C., Schuster, T., Portnoi, T., Barzilay, R., 2019. A deep learning mammography-based model for improved breast cancer risk prediction. Radiology 292, 60–66.

You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B., 2023. Cxr-clip: Toward large scale chest x-ray language-image pre-training, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 101–111.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S., 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, in: Proceedings of the IEEE international conference on computer vision, pp. 19–27.

## Appendix A. Data sampling for image-prompts experiment

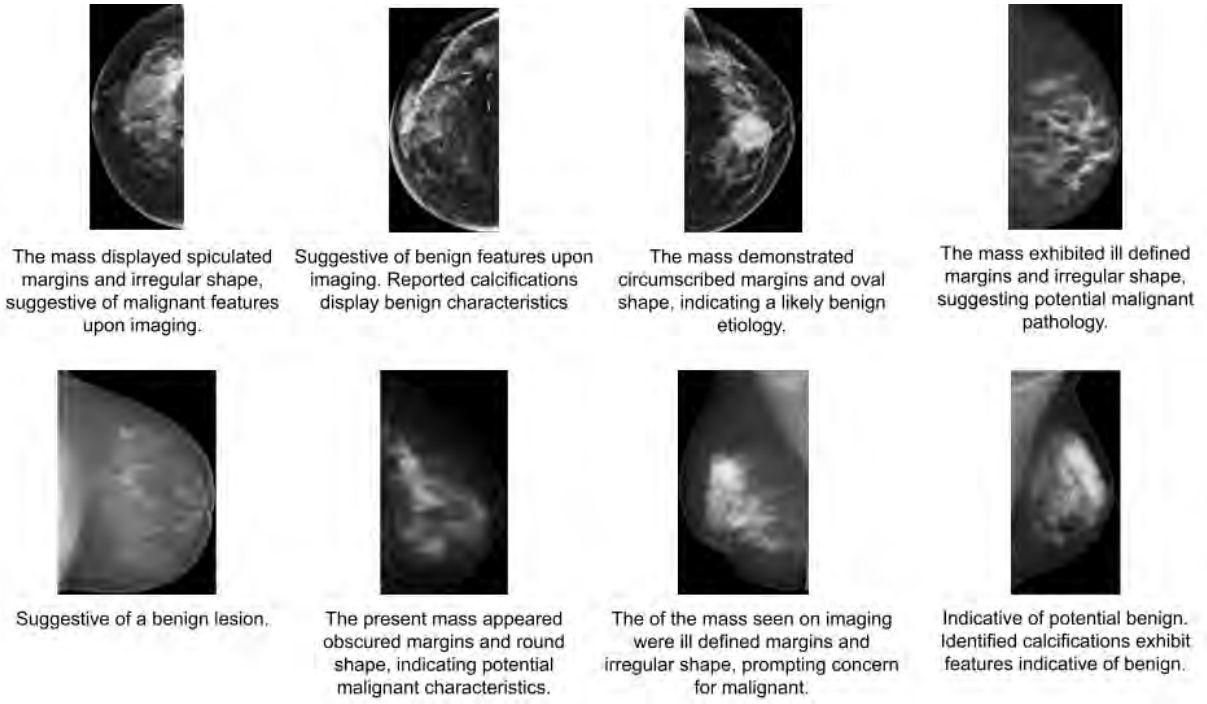



Figure .9: Example of image-prompts pairs sampled from training dataset.

## Appendix B. Prompt generation approach using labelled data.

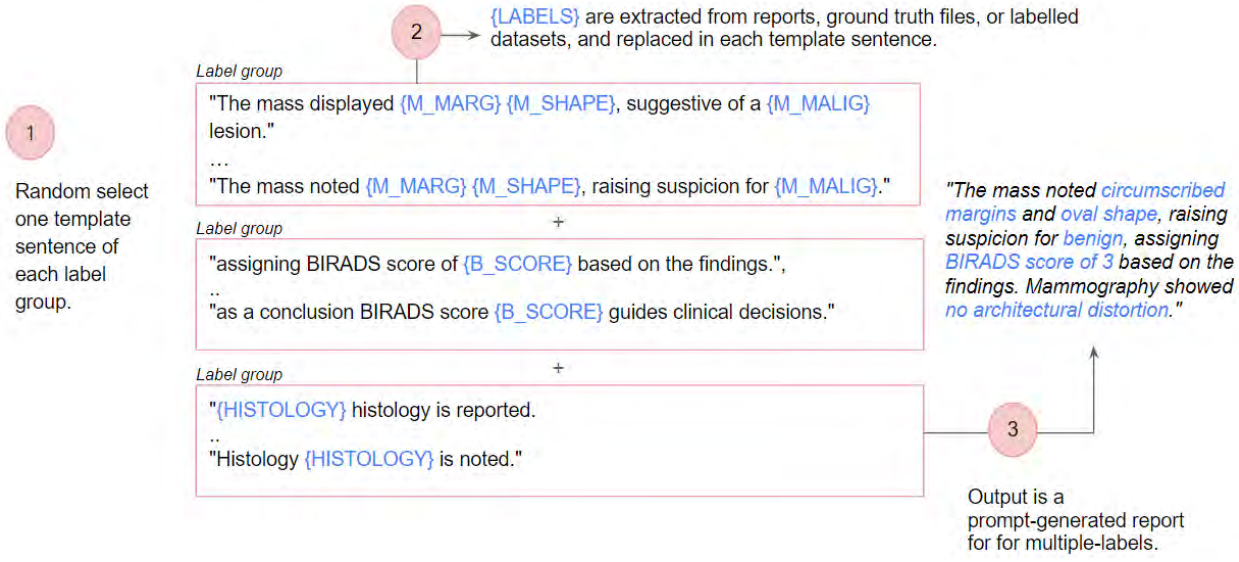
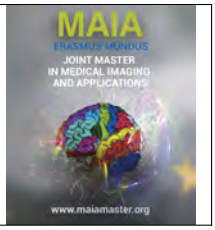



Figure .10: Demonstration of prompt generation mechanism using multiple labels.

Medical Imaging and Applications

Master Thesis, June 2024




# Therapy Response Prediction in Patients with Metastatic Soft Tissue Sarcomas Based on CT Scans Using Delta Radiomics

Frederik Hartmann, Douwe Spaanderman, Martijn Starmans

*The Biomedical Imaging Group Rotterdam, Erasmus MC, the Netherlands*

**Abstract**

Soft tissue sarcomas (STS) are a rare group of cancers that occur in various parts of the body. When metastasized, median survival is only twenty-four months, highlighting the importance of accurate prediction of therapy response to enable the earliest possible modification of treatment. To facilitate this task, we propose a machine learning model that predicts systemic therapy response for metastatic STS based on a pre-treatment and 3-4 months post-treatment CT scan. In particular, the predictions are made using delta features, which represent the longitudinal change of features, with survival time and status as the ground truth. In order to find such a model, we propose an automatic machine learning method for survival analysis, reducing the risk of model selection bias. Furthermore, we investigate various methods for merging features from multiple metastases and compare relative delta features (changes relative to the baseline scan) with absolute delta features (total difference). The evaluation of our method employed nested cross-validation on a training set (n=43), where the best-performing model achieved a concordance index of 0.75±0.14. This model was further tested on a hold-out test set (n=15), achieving a concordance index of 0.70. In this context, it was found that relative delta features improve performance significantly (t-statistic=3.881, p-value=0.0008). Additionally, image-wise merging strategies are shown to be sufficient, simplifying the annotation process. In conclusion, the proposed method demonstrates the feasibility of predicting therapy response for metastatic STS using delta radiomics. To move towards clinical use, it is essential to evaluate the robustness of the presented model using a larger external test cohort, accounting for variations in scanners and acquisition protocols.

*Keywords:* Therapy Response Prediction, Soft Tissue Sarcomas, Automatic Machine Learning, Delta Radiomics

## 1. Introduction

The survival rates for soft tissue sarcoma (STS) present a challenging picture: out of ten patients, only six are expected to be alive after five years (Stiller et al., 2018). Furthermore, for patients with metastatic STS, the survival rate decreases considerably, with the overall five-year survival rate dropping to just one in four, leading to a median survival time of twenty-four months (Lochner et al., 2020). In order to prolong survival, surgery is recommended by the European Society for Medical Oncology (ESMO) as the preferred choice of treatment for patients with metastatic STS (Gronchi et al., 2021). However, in some patients, surgery might not be feasible due to tumor location or size. In such cases, systemic treatment is recommended (Gronchi et al., 2021). If the initial treatment is ineffective, ESMO guidelines suggest switching to a second-line therapy. In addition to changing treatment, discontinuation of treatment is another alternative, as the treatment toxicity may be greater than its predicted benefits (Gronchi et al., 2021). In the best scenario for the patient, the treatment is effective and can be continued as planned. Given that these treatment decisions, such as discontinuation, are based on the patient's response to therapy, it is clear that accurate prediction of therapy response is of upmost importance.

Accurately predicting the response to therapy requires careful consideration of the characteristics of the disease. In particular, STS are a rare group of tumors with more than 80 subgroups. As the name suggests, they are present in soft tissues all around the body, such as the extremities, trunk and retroperitoneum. In

a fifteen-year long study, Coindre et al. (2001) found that 35% of STS patients will develop metastasis. While metastases in the lungs are the most prevalent, it is common for metastases to occur in other sites, such as the liver or lymph nodes (Lochner et al., 2020). This raises a number of difficulties from a machine-learning perspective. First, due to the rarity of the disease, datasets tend to be rather small. Second, the different sites of metastasis reduce the consistency of the images. This increases the difficulty for machine learning, as multiple contrasts between metastases and backgrounds, as well as spatial contexts, have to be considered. Metastases may even be present in several locations in the same scan. To address these difficulties, machine learning methods often perform feature extraction based on segmentations.

The current state of the art in therapy response prediction for STS using machine learning focuses only on the primary tumor (see section 2). The therapy response of the metastatic STS is overlooked. In an attempt to fill this gap, this study is pioneering in the use of machine learning to predict therapy response in metastatic STS.

## 2. State of the Art

In this section, the state of the art in therapy response prediction for STS, both from a clinical and a technical point of view, will be critically discussed.

### *2.1. Clinical*

In clinical practice, the state of the art for predicting therapy response in patients with STS are the RECIST 1.1 guidelines (Eisenhauer et al., 2009). In order to classify the metastasis response, RECIST is based on the relative change of the sum of the longest diameters. The sum consists of the largest axial diameters of multiple metastases. A maximum of five metastases and a maximum of two metastases per organ have to be kept. Based on the relative change of the sum before and after treatment, the treatment response can be defined. However, in high-grade metastatic STS RECIST has been shown to fail in accurately determining treatment response, as noted by Stacchiotti et al. (2009) and Meyer and Seetharam (2019). Given these limitations, RECIST is no longer employed at the Erasmus Medical Center for predicting therapy response in STS in clinical practice, where the present study was conducted. Consequently, clinicians are actively seeking alternative methods.

### *2.2. Technical*

In recent years, machine and deep learning have seen increased usage for STS with a variety of clinical applications. According to Crombé et al. (2023), the works can be divided into four different tasks. First, the discrimination of benign, intermediate and malignant STSs as well as differentiating phenotypes. Second, the prediction of the histological grade. Third, the prediction of survival. Fourth, therapy response prediction for neoadjuvant treatments. As the goal of this work - the therapy response prediction of STS - is most related to category four, the other categories will not be discussed.

One of the earliest studies on predicting therapy response was conducted on patients with histologically confirmed high-grade STS without metastasis Crombé et al. (2019). For the prediction, T1 and T2 weighted Magnetic Resonance Imaging (MRI) scans were used, taken before and after two cycles of chemotherapy. Feature extraction was performed using in-house software based on manually generated tumor segmentations. Additionally, handcrafted categorical features from radiologists were combined with the previously mentioned features. Delta features were calculated as the difference of features between follow-up and baseline. Crombé et al. opted to convert the task of survival prediction to a classification task by assessing the histological response. A good response was defined by having less than 10% of the tumor's cells remaining viable. This allowed to split the patient population into a group of good and bad responders. The best performance of this classification task was achieved using a random forest classifier with an accuracy of 0.75.

Gao et al. (2020) aimed to predict the treatment effect of preoperative radiotherapy in patients with STS of all grades. The ground truth was obtained by comparing the tissue of the primary tumor before and after the surgery. To this end, the relative change of necrotic or fibrous tissue compared to the baseline sample was calculated. A threshold of fifty percent was used to divide the patients into two groups: good responders and bad responders. Features were extracted from diffusion-weighted images using PyRadiomics (van Griethuysen et al., 2017) at three different time points, before radiotherapy, after three sessions and after final radiation treatment. Additionally, delta features were determined by calculating the differences in features between baseline, three cycles and the final treatment. For the classification task, a support vector machine was tested and obtained an area under the curve of $0.90 \pm 0.06$. As the score was significantly lower without the delta features ($0.73 \pm 0.07$.), Gao et al. concluded that delta features were necessary to accurately predict the treatment effect.

Unlike Gao et al., who used two time points to define pathological complete response, Peeken et al. employed a single time point after surgery. A positive treatment response is defined when fewer than five percent of viable cells are necrotic. The features are extracted from T1 and T2 weighted images using PyRadiomics. The delta features were calculated as the difference of features between the two time points. For the classification task of predicting the pathological complete response, a nested cross-validation is used to compare be-

tween Random Forest, ElasticNet Regression and LogitBoost. Overall, Random Forest performs best using delta-features, resulting in an area under the curve (AUC) of 0.75 on an external validation set. It is important to highlight that the result is only marginally worse than the result on the internal training set using nested cross-validation with an AUC of 0.73. Moreover, Peeken et al. pointed out that this classification task is a substitute task for predicting overall survival. Therefore, the output of the Random Forest was employed as a feature for predicting overall survival using a multivariate Cox proportional hazards model, resulting in harrell's concordance index (c-index) of 0.69.

Fields et al. (2023) claim that the previously described methods are not reproducible due to the feature selection techniques used. Similar to the methods described above, good and bad treatment responses are defined based on pathological findings; however, the specifics are not detailed. In contrast to the previously described methods, in-house MATLAB software was utilized for feature extraction. The extraction was performed on eleven different MRI sequences. Delta features were calculated as the difference between pre- and post surgical features. Afterwards, Random Forest and AdaBoost were used for the classification task on the full feature set, resulting in an AUC of 0.44 and 0.40 respectively.

Following Peeken et al., Miao et al. (2023) convert the task into a classification task by using the pathological complete response to divide patients into groups of good and bad responders. In comparison to Gao et al. and Peeken et al., the authors extend the inclusion criteria to a broader variety of therapies. The features were extracted from T1, T2 and diffusion-weighted images using PyRadiomics. Delta features were calculated as the difference between pre- and post-surgery time points. Using multivariate logistic regression, the authors achieve an AUC of 0.952.

In summary, the papers presented demonstrate the possibility of predicting therapy response in STS from a variety of different MRI sequences, such as T1, T2 or diffusion weighted imaging. Gao et al. showed the advantage of delta features compared to single-time points and both Peeken et al. as well as Miao et al. were able to reproduce this on their respective datasets. Furthermore, Miao et al. explored the feasibility of extending the models to different treatment types.

### 2.3. Delta Radiomics for Therapy Response Prediction

The work described in subsection 2.2 differs from this thesis in three ways regarding patient data. First, for this study, Computed Tomography (CT) scans are used instead of MRI. Second, multiple metastasis segmentations per scan are present. Third, the prediction is done using survival data instead of pathological ground truth. In order to provide a picture of the state of the art with respect to these differences, three additional papers not explicitly related to STS will be discussed.

The first work by Nardone et al. (2020) evaluates the robustness of texture features on different CT scanners. For this, a variety of tissues, including those of the lung and liver, were tested in a phantom study. Three scanners were used for evaluation, with four protocols for each scanner. The study showed that delta features are more robust among different scanners and protocols compared to features from a single time point.

Making use of the robustness of delta features in CT scans, Qu et al. (2023) predicted therapy response in metastatic colorectal liver cancer. Instead of using the histological response, RECIST was used as a ground truth to classify between a complete and partial response. The features were extracted from contrast-enhanced CTs using PyRadiomics from pre- and post-therapy scans. Qu et al. showed that relative delta features, such as relative tumor growth, performed better than absolute delta features, such as absolute volume increase. It is worth pointing out that the features were extracted using segmentations of the biggest metastasis.

On the other hand, Cousin et al. (2023) extracted features from multiple segmentations per CT scan with the goal of predicting therapy response for advanced non-small-cell lung cancer. The segmentations were carried out following RECIST guidelines, but instead of using RECIST to define the ground truth as well, the survival data was used directly. Comparing the effect of multiple segmentations versus a single one, Cousin et al. points out that concatenation of features from multiple lesions leads to a worse performance compared to a single segmentation. Nevertheless, for both, the feasibility of therapy response prediction using survival data was shown.

### 2.4. Beyond the State of the Art

In addition to the differences in imaging modality, this study aims at five key improvements to address the limitations of the state of the art. First and foremost, the clinical goal is different from the papers presented in subsection 2.2. While Crombé et al.; Fields et al.; Gao et al.; Peeken et al. and Miao et al. aim to predict therapy response for the primary tumor, this work aims to predict therapy response for metastases in STS. Second, the ground truth is not based on histological findings, as in previous works, but on survival data. It is worth mentioning that Peeken et al. tried to accomplish this by using the logits of the classification as an input for survival analysis. In comparison, this work directly predicts therapy response using survival data. This is because histologic tissue samples for metastases are uncommon, as this would require multiple biopsies for each patient, which are not performed due to the clinical risks associated with the invasive nature of such procedures. Furthermore, Peeken et al. note that the

classification task was only used as a surrogate for survival prediction. The third difference compared to previous works on STS is the use of relative delta features instead of absolute ones. The advantage for this was shown by Qu et al. for colorectal liver metastases, but not yet for therapy response prediction in patients with STS. Fourth, segmentations of multiple metastases are available for each scan. As Cousin et al. pointed out, in the case of non-small-cell lung cancer, the concatenation of feature spaces was insufficient. Thus, this work tries to overcome this by aggregating the features from multiple segmentations by means of, e.g., summation or averaging. Finally, it is worth pointing out that due to the fact that texture delta features are more robust than single time point features (Nardone et al., 2020) and the fact that several authors such as Miao et al. or Gao et al. have already shown the advantage of delta features for STS, single time point prediction is not performed in this study.

In summary, the key contributions of this work are as follows: First, this work pioneers therapy response prediction for metastatic STS through machine learning methods. Second, evaluating the impact of absolute and relative delta features in STS patients. Third, merging strategies for features extracted from multiple metastases in patients with STS are proposed and assessed.

## 3. Material & Methods

### *3.1. Data*

In this section, the data collection and segmentation processes will be explained. Furthermore, the feature extraction, merging and selection methods are presented.

**The Dataset in One Look**

| | |
|---|---|
| **Collected from** | Erasmus MC |
| **Modality** | Contrast-enhanced CT |
| **Scan region** | Trunk |
| **Number of Patients** | 58 |
| **Scans per Patient** | 1 pre-, 1 post-treatment |
| **Segmentation 1** | InteractiveNet |
| **Segmentation 2** | Manual adjustment |
| **Ground Truth** | Survival status & time |

Figure 1: The dataset summarized. Metastases from 58 STS patients treated at the Erasmus Medical Center in Rotterdam, the Netherlands, were segmented in contrast-enhanced CT scans of the trunk using InteractiveNet, followed by manual adjustment. Each patient had one pre-treatment and one post-treatment scan, with ground truth including survival status and time.

#### *3.1.1. Patient Selection*

The Erasmus Medical Center is an expertise center for STS which treats most patients with STS in the region, including patients with metastatic disease. The data of patients with STS who were treated at the EMC were collected retrospectively. This study was conducted in accordance with the ethical standards outlined in the Helsinki Declaration of 1975 and was exempted by the medical research ethics committee (METC, MEC-2020-0687) from the Erasmus Medical Center with a waiver of consent. The inclusion criteria of the study are 1) the patient has been diagnosed with an initially unresectable, metastasized, high-grade STS, 2) the patient has received systemic treatment at the Erasmus Medical Center between 2014 and 2020. Patients were excluded based on the following criteria; 1) the patient has gastrointestinal stromal tumors, 2) the patient has no detectable or suitable metastasis during follow-up or no follow-up after a maximum of four months of treatment, 3) the patient has necrotic or cystic metastasis 4) CT quality too low for measuring response criteria as determined by a musculoskeletal radiologist with 7 years of experience.

The patient selection process can be seen in Figure 2 and the characteristics of the selected patients in Table 1. For all selected patients, CT scans have been taken as part of the systemic treatment. These include a contrast-enhanced CT scan of the trunk taken before the start of the systemic treatment. Each scan shows at least one metastatic location. The primary tumor might not be present in the scan. Additionally, at least one follow-up scan was taken not later than four months after the start of the treatment, in accordance with the ESMO guidelines.

#### *3.1.2. Segmentation*

The contrast-enhanced CT scans of 58 patients were segmented at both the baseline and the first follow-up. While CT scans of 37 additional patients were available, the segmentations could not be carried out due to time constraints. The segmentation followed RECIST 1.1 guidelines. In summary, only the five biggest metastases were segmented, with a maximum of two metastases per organ. Two segmentations were used to compare the feature robustness to different segmentations. First, segmentations were created using InteractiveNet Segmentation by Spaanderman et al. (2024). The deep learning-based method requires the selection of six points near the tumor's extreme boundaries. These points are then transformed into an exponentialized geodesic map. The map is then combined with the original image to predict the segmentation using a 3D U-net architecture. The second segmentation is the adjustment of the results of the first method. The adjustment was carried out manually for each segmentation. The refinement was done to improve on potential shortcomings of the prediction. Three differences in the segmentations are displayed in

Table 1: Characteristics of the included patients. Categories with fewer than two patients are consolidated into "Others." Patients with unknown characteristics are excluded from the listing but are accounted for in the percentage calculations.

| **Characteristics** | **train** (n=43) | **test** (n=15) |
|---|---|---|
| **Age** years | | |
| Median [IQR] | 54 [47; 66] | 57 [43; 65] |
| Mean | 54 | 52 |
| **Sex** n(%) | | |
| Male | 21 (49%) | 10 (67%) |
| Female | 23 (51%) | 5 (33%) |
| **Survival status** n(%) | | |
| Dead | 34 (79%) | 10 (67%) |
| Loss of Follow-up | 1 (2%) | 1 (7%) |
| Alive | 9 (19%) | 4 (26%) |
| **Survival time** months | | |
| Median [IQR] | 22 [14; 42] | 23 [10; 58] |
| Mean | 35.14 | 47.35 |
| **Overall survival** % | | |
| Three years | 43 | 37 |
| Five years | 21 | 37 |
| **Phenotype** n(%) | | |
| Leiomyosarcoma | 15 (34%) | 4 (26%) |
| Pleomorphic sarcoma | 7 (16%) | 2 (13%) |
| Angiosarcoma | 5 (11%) | 1 (7%) |
| Liposarcoma | 4 (9%) | 2 (13%) |
| Synovial sarcoma | 3 (7%) | 0 (0%) |
| Others | 10 (23%) | 6 (40%) |
| **Site of Primary Tumor** n(%) | | |
| Leg | 8 (19%) | 3 (20%) |
| Uterus | 8 (19%) | 2 (13%) |
| Abdomen | 2 (5%) | 2 (13%) |
| Arm | 2 (5%) | 1 (7%) |
| Breast | 2 (5%) | 0 (0%) |
| Sacroiliac Joint | 2 (5%) | 0 (0%) |
| Scalp | 2 (5%) | 1 (7%) |
| Shoulder | 2 (5%) | 1 (7%) |
| Retroperitoneal | 0 (0%) | 2 (13%) |
| Other | 14 (32%) | 5 (33%) |
| **Site of Metastasis** n(%) | | |
| Lung | 21 (49%) | 5 (33%) |
| Multiple sites | 6 (14%) | 7 (46%) |
| Lymph nodes | 3 (7%) | 0 (0%) |
| Retroperitoneal | 3 (7%) | 0 (0%) |
| Other | 10 (23%) | 2 (13%) |
| **Time to Metastasis** n(%) | | |
| Metachronous | 25 (57%) | 9 (60%) |
| Synchronous | 19 (43%) | 5 (33%) |

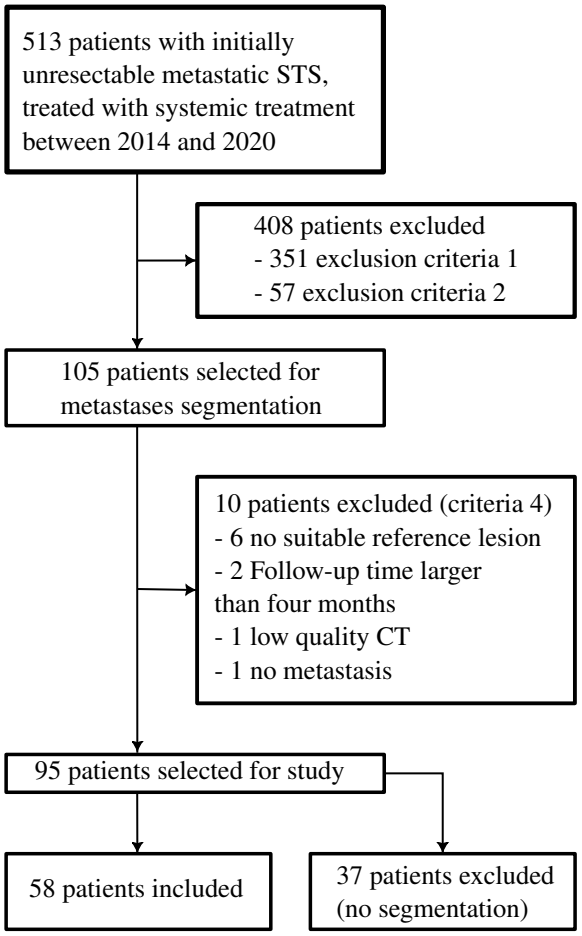



Figure 2: Flowchart of the patient selection included in this study.

Figure 3. Both segmentations were carried out by a medical student under the supervision of a radiologist with eight years of experience.

### *3.1.3. Ground Truth*

The ground truth is based on the survival status of the patient at the time of this study. If a patient was alive at the time of data collection (the 30th of September, 2023), the survival time was calculated as the absolute difference between the date of data collection and the date of STS diagnosis. If the patient was dead or censored, the survival time was calculated as the absolute difference between the date of death or date of the last follow-up and the date of diagnosis.

### *3.1.4. Feature Extraction*

Following the segmentation, feature extraction was performed. To allow for comparability, an open-source feature extraction library named PyRadiomics (van Griethuysen et al., 2017) was chosen. PyRadiomics takes as input the scan and the segmentation of the volume of interest. In STS patients with metastatic disease, multiple metastases are present in each scan. However, PyRadiomics only allows for the feature extraction of a sin-

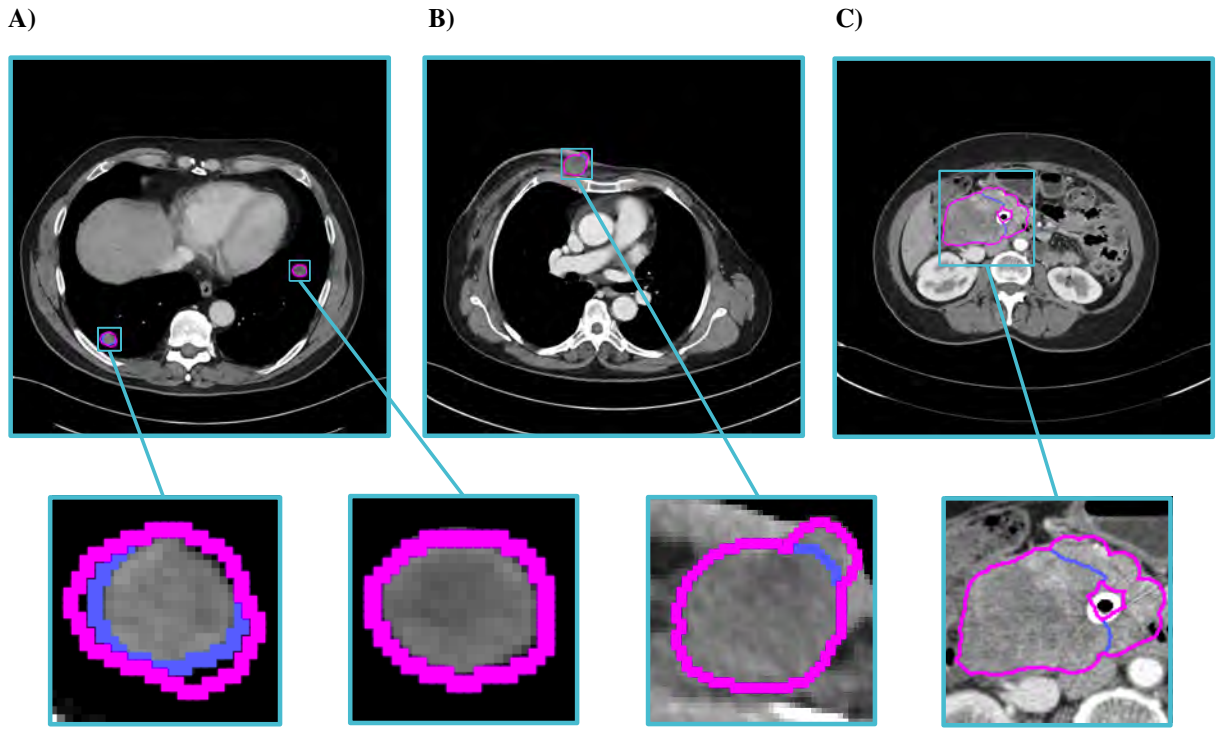



Figure 3: Segmentation results are presented from InteractiveNet (blue outline) and manually refined segmentation (magenta outline). Figures A, B and C illustrate metastasis segmentations, respectively, showing the minimum, median (711 voxels), and maximum (24169 voxels) differences of the two segmentations within the dataset. Figure A illustrates a lung metastasis; Figure B shows a subcutaneous or bone metastasis; and Figure C shows a pancreatic metastasis.

gle label or metastasis. To take advantage of the information present in each metastasis, a feature set was extracted for every segmented metastasis. Therefore, multiple feature sets were extracted per scan. Each feature set consists of the same 108 features. The features are the 107 PyRadiomics features with default settings (version 3.1.0) and the number of segmented lesions per scan. The feature categories can be seen in Table 2.

### *3.1.5. Delta features*

The prediction of the therapy response can be interpreted as the monitoring of changes in features over time. This longitudinal change in features is referred to as delta features. A mathematical definition is given in Equation 1. Let $x^t$ denote a feature at timepoint $t$. The difference, expressed as $d$, can be defined as

$$d \stackrel{\text{def}}{=} x^{Follow-up} - x^{Baseline} \tag{1}$$

Another option is to take the relative difference as displayed in Equation 2.

$$d_{rel} \stackrel{\text{def}}{=} \frac{x^{Follow-up} - x^{Baseline}}{x^{Baseline}} \tag{2}$$

The equations above assume one feature value for each feature. As multiple metastases are present in each scan, multiple feature values are extracted for each feature. This requires the adaptation of the equations displayed above. One option would be to allow multiple values per feature and to concatenate them. However, this could create excessively large feature spaces of varying sizes. In order to maintain a consistent size of the feature spaces, three different merging strategies are explored. These techniques are the arithmetic mean, the maximum and the summation. Furthermore, the

Table 2: Categories of PyRadiomics Features

| Feature Category | Number of Features |
|---|---|
| First-order features | 19 |
| Shape-based features | 16 |
| Features from the gray-level co-occurrence matrix (glcm) | 24 |
| Features from the gray-level run-length matrix (glrlm) | 16 |
| Features from the gray-level size-zone matrix (glszm) | 16 |
| Features from the neighboring-gray-tone difference matrix (ngtdm) | 5 |
| Features from the gray-level dependence matrix (gldm) | 14 |

merging can be performed at different points. The first option is to extract the features for each metastasis, aggregate them (e.g., by summation) and calculate the absolute (Equation 1) or relative difference (Equation 2) afterwards. In comparison to that, the second option starts by calculating the difference and aggregates the features afterwards. For this option, the differences are calculated for each metastasis pair, i.e., the same metastasis at baseline and follow-up. If the matching metastasis is not present, it is replaced by a feature vector of zeros. A reason for a missing match could be a full remission of the metastasis in the follow up. If the matching metastasis is missing in the baseline scan, e.g., due to the growth of a new metastasis, this would lead to a division by zero in Equation 2. In this case, the relative features are set to one. The second option can be seen in Figure 4. In the following, all strategies are presented in detail. First, all three equations for metastasis-wise merging are presented. Afterwards, the three equations for image-wise merging are shown. For each equation, the calculations for absolute and relative delta features are shown.

**Metastasis-wise Merging**

The metastasis-wise delta features merged through the mean can be calculated with Equation 3.

$$\Delta x_{mets} = \frac{1}{n}\sum_{i=1}^{n} d_i \quad \text{or} \quad \Delta x_{rel,mets} = \frac{1}{n}\sum_{i=1}^{n} d_{rel,i} \tag{3}$$

Where $n$ is the number of segmented metastases and $i$ represents a metastasis pair. The calculation for metastasis-wise merging with the maximum is shown in Equation 4.

$$\Delta x_{mets} = \max_{i=1}^{n} d_i \quad \text{or} \quad \Delta x_{rel,mets} = \max_{i=1}^{n} d_{rel,i} \tag{4}$$

Finally, summation aggregates the total value of the differences:

$$\Delta x_{mets} = \sum_{i=1}^{n} d_i \quad \text{or} \quad \Delta x_{rel,mets} = \sum_{i=1}^{n} d_{rel,i} \tag{5}$$

**Image-wise Merging**

Equation 6 describes the image-wise merging using the arithmetic mean.

$$\Delta x_{img} = \frac{1}{n}\sum_{k=1}^{n} x_k^{Follow-up} - \frac{1}{m}\sum_{k=1}^{m} x_k^{Baseline} \quad \text{or}$$
$$\Delta x_{rel,img} = \frac{\frac{1}{n}\sum_{k=1}^{n} x_k^{Follow-up} - \frac{1}{m}\sum_{k=1}^{m} x_k^{Baseline}}{\frac{1}{m}\sum_{k=1}^{m} x_k^{Baseline}} \tag{6}$$

$n$ represents the number of segmented metastases at baseline and $m$ at follow-up. Furthermore, $k$ does not indicate a matching metastasis pair anymore but is a simple iterator instead. Similarly, the image-wise merging with the max is shown in Equation 7.

$$\Delta x_{img} = \max_{k=1}^{n} x_k^{Follow-up} - \max_{k=1}^{m} x_k^{Baseline} \quad \text{or}$$
$$\Delta x_{rel,img} = \frac{\max_{k=1}^{n} x_k^{Follow-up} - \max_{k=1}^{m} x_k^{Baseline}}{\max_{k=1}^{m} x_k^{Baseline}} \tag{7}$$

Summing it all up, the image-wise merging using the sum is presented in Equation 8

$$\Delta x_{img} = \sum_{k=1}^{n} x_k^{Follow-up} - \sum_{k=1}^{m} x_k^{Baseline} \quad \text{or}$$
$$\Delta x_{rel,img} = \frac{\sum_{k=1}^{n} x_k^{Follow-up} - \sum_{k=1}^{m} x_k^{Baseline}}{\sum_{k=1}^{m} x_k^{Baseline}} \tag{8}$$

Each of the six merging strategies is applied to the entire collection of images, resulting in six initial datasets. By adding both relative and absolute options and using two different segmentations, the sets are further diversified. This results in a total of twenty-four distinct datasets, each produced by a unique combination of strategies, options and segmentations. For each dataset, the same patients have been split into a training and test set with a split of 75% vs 25%.

*3.1.6. Feature Selection*

For each training set, three feature selection methods are explored. 1) variance threshold and recursive feature elimination with cross-validation; 2) variance threshold and repeated recursive feature elimination with cross-validation; 3) no feature selection.

Both the first and second method employ a variance threshold. The variance threshold removes features that have a lower variance than the defined threshold. Here, following the manual of scikit-learn, whose implementation was used (Pedregosa et al., 2011), a threshold of 0.8 is chosen. The second step, recursive feature elimination with cross-validation, is only available for regression and classification models in scikit-learn but not for survival models. To test recursive feature elimination with cross-validation, a custom version based on recursive feature elimination (without cross-validation) of scikit-learn is implemented. The advantage of using the version with cross-validation is the automatic selection of the number of features, which can reduce selection bias. Recursive feature elimination with cross-validation aims to overcome this by trying every possible number of features to be selected. This is done by performing recursive feature elimination in a cross-validation setup for each number. For each number, the models are evaluated based on the mean metric of the cross-validation setup. Afterwards, the number of features with the highest mean metric is chosen. This number is used to perform recursive feature elimination, but

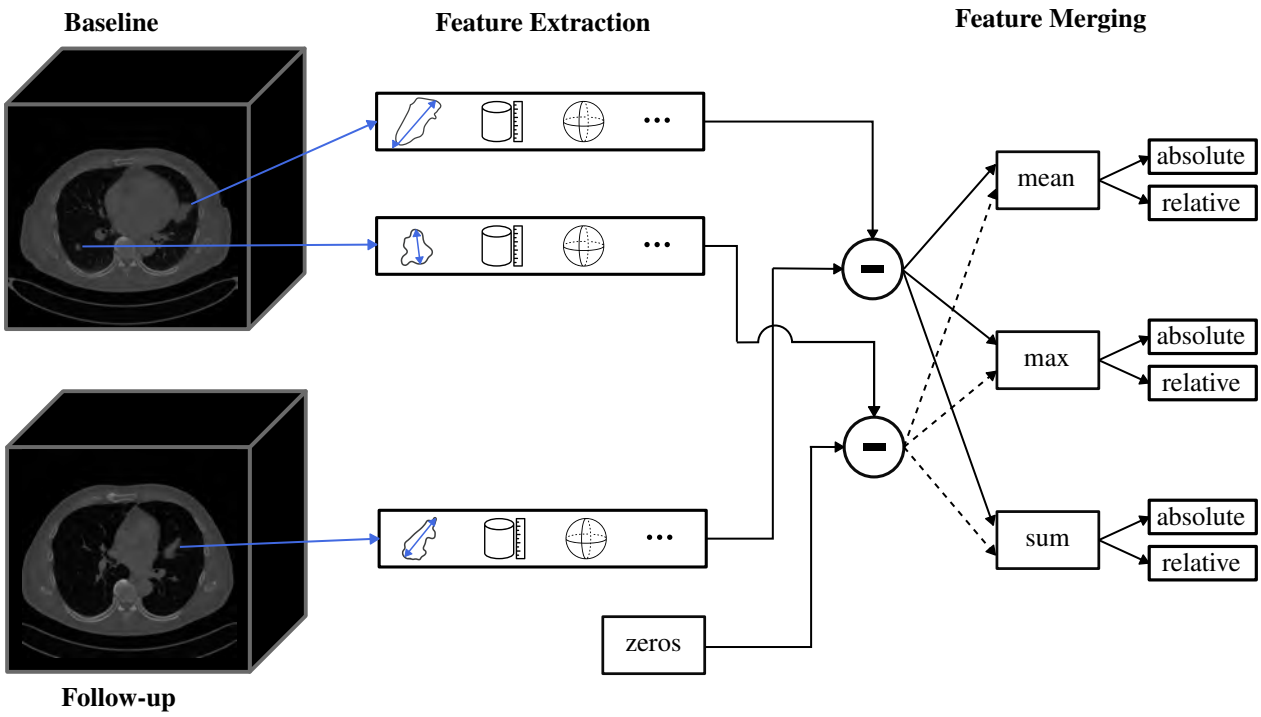



Figure 4: Feature merging is performed for each metastasis. First, for every segmented metastasis, features are extracted using PyRadiomics. The match for the second metastasis is not present in the follow-up anymore and is therefore replaced by a feature vector of zeros. Afterwards, the difference for every metastasis pair is calculated. The differences are then merged using either the arithmetic mean, max or summation. Finally, both the absolute and relative delta features are calculated. This results in six different delta feature sets per patient.

this time on the entire training set. The output of this training setup is the output of recursive feature elimination with cross-validation. Here, a five-fold cross-validation is chosen with the c-index as a metric. As a model for the recursive feature elimination, GradientBoostingSurvivalAnalysis by scikit-survival (Pölsterl, 2020) is used. This choice is mainly based on the fact that the model fulfills all requirements of the recursive feature elimination, mainly the feature importance's. The step size of the recursive feature elimination is set to one, following the implementation in scikit-learn.

### 3.2. *Automatic Machine Learning*

For the final prediction of the therapy response, an auto-ML method is proposed. This method aims to select the best-performing machine learning pipeline automatically while simultaneously reducing the risks of overfitting. The selected pipeline consists of three steps. First, outlier detection & imputation. Second, feature scaling and third, the survival model. Every possible combination of methods is tested and evaluated while maintaining the order of the pipeline. Furthermore, a skipping step for the first two pipeline steps is implemented. To reduce the risk of model selection bias, this is done in a nested cross-validation setting. This pipeline selection process is carried out separately for every training set and feature selection method. Combining twenty-four datasets and three feature selection methods results in seventy-two pipelines. Of these pipelines, the best-performing model is selected and evaluated on the independent test set. Only the best model is tested to avoid a selection bias and hence reduce the chance of overfitting even further. An example of the pipeline selection process is shown in Figure 5.

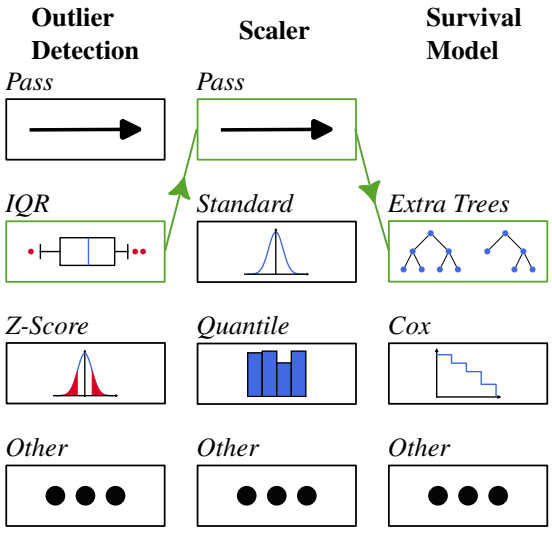



Figure 5: The pipeline selection search space consists of one from each category: outlier detection & imputation, scaler and survival model. One exemplary pipeline consisting of interquartile range detection, no feature scaling and extra trees as a survival model is highlighted in green.

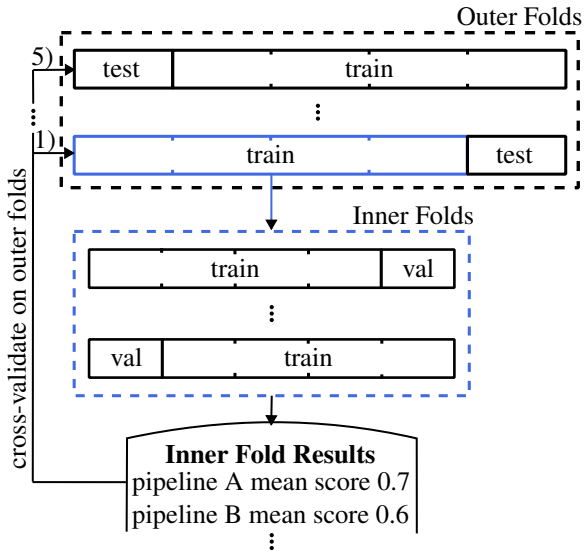



Figure 6: Workflow of nested cross validation.

### 3.2.1. *Nested Cross-Validation*

In the following, the nested cross-validation used for the pipeline selection process is explained in detail. First, the five-fold cross-validation splits into outer test and train folds. Second, inner five-fold cross-validation splits the outer training fold into inner validation and training folds. For each inner fold, a search for the best pipeline is carried out. By averaging the results among the inner validation folds, a performance measure for each pipeline can be found. By selecting the best pipeline from each inner fold, five pipelines remain. These five pipelines are then trained and evaluated on the outer folds. The best-performing pipeline is the one with the highest average score on the outer test folds. The workflow is shown in Figure 6.

### 3.2.2. *Outlier Detection & Imputation*

The first step of the pipeline is outlier detection and imputation. In total, three different methods are implemented. The first method is based on the interquartile range. Using the interquartile range, a value is defined as an outlier if it is bigger or smaller than the closest whisker. The outliers are then replaced with the closest whisker. The second method is based on the z-score. For this, the feature values are z-scored as described in Equation 9.

$$x_z = \frac{x - \mu}{\sigma} \tag{9}$$

Afterwards, an outlier is present if $x_z <$ threshold. The threshold is set at three, following the example of Hoaglin and Iglewicz (1993). The outliers are then set to the threshold. The third outlier detection technique is based on the modified z-score after Hoaglin and Iglewicz (1993).

$$x_{mod} = \frac{0.6745 \cdot (x - \text{median}(x))}{\text{MAD}} \tag{10}$$

where MAD is the median absolute deviation. The threshold here is set to 3.5 as recommended by Hoaglin and Iglewicz (1993).

$$\text{MAD} = \text{median}\left(|x - \text{median}(x)|\right) \tag{11}$$

### 3.2.3. *Feature scaling & Survival Models*

Steps two and three of the pipeline are the scaling of features and finally, the prediction of the therapy response using a survival model. The feature scalers of Scikit-learn are used. The feature scalers include, MaxAbs-Scaler, Min-Max-Scaler, Robust-Scaler, Standard-Scaler and Quantile-Transformer with 25 quantiles. The survival models employ the implantation of scikit-survival and are Survival Tree, Coxnet Survival Analysis, Componentwise Gradient Boosting Survival Analysis, Gradient Boosting Survival Analysis, Random Survival Forest and Extra Survival Trees.

## 3.3. *Robustness & Interpretability*

The best pipeline, in combination with its corresponding segmentation & merging strategy, is chosen based on the highest mean c-index across the outer folds. Once the best pipeline is found, the pipeline is retrained on the full training set and the performance on the independent test set is computed. To evaluate the pipeline and selected features under varying conditions, four additional tests are conducted. The first three tests involve retraining the model with different modifications, such as altering segmentation methods, using absolute instead of absolute delta features, and applying either metastasis or image-wise merging options. The fourth test aims to evaluate how the model would perform in a clinical setting. There, the clinician segmenting a new image would be a different one than the one who annotated the training images. In an attempt to recreate this scenario, the model is trained on one segmentation and tested on the other.

To enhance the interpretability of the best pipeline, Survshap (Krzyziński et al., 2023) is employed. The SHAP (SHapley Additive exPlanations) values allow for the identification of how each feature contributes to each individual prediction, quantifying the impact in a consistent and accurate manner. They represent the marginal contribution of each feature to the prediction outcome, thereby providing insights into the decision-making process of the model. This facilitates a better understanding of the model's behavior and supports transparent and explainable predictions in survival analysis.

### 3.4. Metrics

The concordance index is computed as as:

$$C = \frac{\sum_{i<j} \mathbf{1}(T_i < T_j) \cdot \mathbf{1}(\hat{T}_i < \hat{T}_j) \cdot \delta_i}{\sum_{i<j} \mathbf{1}(T_i < T_j) \cdot \delta_i} \tag{12}$$

where:

- $T_i$ and $T_j$ are the actual survival times of patients $i$ and $j$
- $\hat{T}_i$ and $\hat{T}_j$ are the predicted survival times for patients $i$ and $j$
- $\mathbf{1}(\cdot)$ is the indicator function, which is 1 if the condition inside is true and 0 otherwise.
- The sum $\sum_{i<j}$ is taken over all pairs of subjects $i$ and $j$ such that $i < j$, which means each pair is considered only once.
- $\delta_i$ is a binary event indicator. 1 corresponds to the death of the patient and 0 being censoring.

In other words, the numerator of the C-index counts the number of correctly ordered pairs of patients, where the patient with the shorter survival time also has a shorter predicted survival time. The denominator counts all possible pairs. The concordance index is computed using scikit-survival Pölsterl (2020). A c-index of 1 resembles a perfect prediction, while a c-index of 0.5 corresponds to a random prediction.

## 4. Results

This section presents the nested cross-validation outcomes, followed by the evaluation of the final model on the test set. Finally, the SHAP analysis is introduced.

### 4.1. Evaluation of datasets

The best-performing pipeline without feature selection achieved a mean c-index of 0.68 with a standard deviation of 0.19, as illustrated in Table 3. The dataset used was built by averaging the metastasis-wise delta features extracted from the manually adjusted segmentations. The pipeline selected by the automatic machine learning was a z-score-outlier transformer, followed by a quantile transformer for feature scaling, and finally extra survival trees for survival prediction. On average, 132.2 feature values were selected as outliers out of a total of 4644 values.

Next, the use of variance thresholding followed by recursive feature elimination with cross-validation resulted in the best overall performance with a c-index of $0.75 \pm 0.14$ in the nested cross-validation. This was achieved by summing the image-wise relative delta features extracted from the adjusted segmentation. The pipeline selected by the auto-ML methods was an interquartile range-based outlier detection, a quantile transformer for scaling, and a random survival forest for therapy response prediction. This methodology performed best overall and was therefore selected as the final model. An average of 45.8 values were detected as outliers from a selected subset with a total of 387 values. The feature space consisted of ten features.

Finally, the best-performing pipeline using variance thresholding followed by repeated recursive feature elimination with cross-validation achieved a c-index of 0.73 with a standard deviation of 0.09. This dataset was created by summing the relative metastasis-wise delta features from the InteractiveNet segmentation. The performance was based on the best-performing pipeline selected by the auto-ML method, which is a Z-score outlier detector followed by a quantile transformer and a Cox net survival analysis model. An average of 9.6 feature values were detected as outliers from a total of 258 values. The six selected features are shown in Table 4. Using the same dataset but employing variance thresholding followed by a single repetition of feature elimination resulted in a performance of $0.7457 \pm 0.1281$, which was the second-best method overall.

#### 4.1.1. Evaluation of Feature Selection Methods

Comparing the three methods using the F-statistic showed a significant difference, with an F-statistic of 23.227 and a p-value of $1.94 \cdot 10^{-8}$. As the F-statistic does not indicate which of these methods caused the difference, the selections were compared against each other. Repeated recursive feature elimination with cross-validation resulted in the best average performance across all datasets, with a mean c-index of 0.691 and a standard deviation of 0.029. This performance was superior to a single repetition, which had a mean c-index of 0.676 and a standard deviation of 0.047. However, considering a t-statistic of $-1.297$ and a p-value of 0.201, the difference between these two methods is not statistically significant.

The comparison of both of these methods to no feature selection revealed that both feature selection methods resulted in a significantly better performance compared to no feature selection, which on average had a c-index of 0.612 with a standard deviation of 0.049. The comparison of no selection versus variance thresholding and a single repetition had a t-statistic of $-4.627$ and a p-value of $3.04 \cdot 10^{-5}$. The comparison of no selection versus variance thresholding and repeated recursive feature elimination with cross-validation had a t-statistic of $-6.784$ and a p-value of $1.92 \cdot 10^{-8}$.

In order to provide further insights into the selected features, Table 4 shows the feature set corresponding to the best-performing dataset of the respective feature selection methods. While most of the selected features are different, two of the selected features were similar. Namely, the shape feature Mesh-Volume and firstorder uniformity. These two features are not identical, as one results from image-wise merging and the

Table 3: Summary of results of nested cross-validation for multiple feature merging and selection strategies. The c-index is reported using the mean and standard deviation of the best pipeline for each strategy. For each pipeline, the performance is calculated among all outer folds.

| | | | | **c-index mean (std)** | | |
|---|---|---|---|---|---|---|
| Segmentation | Aggregation | Abs or Rel | Merged at | No selection | Var and RFECV | Var and repeated RFECV |
| Adjusted | Max | Absolute | Image | 0.63 (0.15) | 0.61 (0.07) | 0.66 (0.15) |
| Adjusted | Max | Absolute | Tumor | 0.65 (0.11) | 0.70 (0.10) | 0.73 (0.15) |
| Adjusted | Max | Relative | Image | 0.56 (0.13) | 0.67 (0.08) | 0.68 (0.14) |
| Adjusted | Max | Relative | Tumor | 0.66 (0.10) | 0.73 (0.06) | 0.69 (0.11) |
| Adjusted | Mean | Absolute | Image | 0.59 (0.07) | 0.64 (0.11) | 0.64 (0.16) |
| Adjusted | Mean | Absolute | Tumor | 0.61 (0.11) | 0.60 (0.11) | 0.67 (0.13) |
| Adjusted | Mean | Relative | Image | 0.67 (0.16) | 0.70 (0.15) | 0.70 (0.15) |
| Adjusted | Mean | Relative | Tumor | **0.68 (0.19)** | 0.67 (0.11) | 0.67 (0.07) |
| Adjusted | Sum | Absolute | Image | 0.62 (0.26) | 0.63 (0.18) | 0.70 (0.09) |
| Adjusted | Sum | Absolute | Tumor | 0.59 (0.23) | 0.71 (0.11) | 0.68 (0.16) |
| Adjusted | Sum | Relative | Image | 0.65 (0.17) | **0.75 (0.14)** | 0.68 (0.09) |
| Adjusted | Sum | Relative | Tumor | 0.69 (0.17) | 0.69 (0.14) | 0.69 (0.12) |
| InteractiveNet | Max | Absolute | Image | 0.56 (0.23) | 0.65 (0.12) | 0.63 (0.16) |
| InteractiveNet | Max | Absolute | Tumor | 0.58 (0.20) | 0.58 (0.14) | 0.69 (0.06) |
| InteractiveNet | Max | Relative | Image | 0.59 (0.08) | 0.73 (0.13) | 0.73 (0.13) |
| InteractiveNet | Max | Relative | Tumor | 0.64 (0.15) | 0.67 (0.08) | 0.72 (0.15) |
| InteractiveNet | Mean | Absolute | Image | 0.52 (0.10) | 0.71 (0.13) | 0.71 (0.13) |
| InteractiveNet | Mean | Absolute | Tumor | 0.62 (0.15) | 0.63 (0.07) | 0.70 (0.12) |
| InteractiveNet | Mean | Relative | Image | 0.67 (0.20) | 0.71 (0.13) | 0.75 (0.13) |
| InteractiveNet | Mean | Relative | Tumor | 0.58 (0.11) | 0.72 (0.05) | 0.70 (0.14) |
| InteractiveNet | Sum | Absolute | Image | 0.57 (0.14) | 0.64 (0.06) | 0.67 (0.08) |
| InteractiveNet | Sum | Absolute | Tumor | 0.54 (0.20) | 0.66 (0.12) | 0.67 (0.08) |
| InteractiveNet | Sum | Relative | Image | 0.67 (0.22) | 0.68 (0.09) | 0.69 (0.08) |
| InteractiveNet | Sum | Relative | Tumor | 0.55 (0.09) | 0.75 (0.13) | **0.73 (0.09)** |

other from metastasis-wise merging. Nevertheless, both are summed relative delta features.

#### *4.1.2. Evaluation of Aggregation strategies*

Three different aggregation strategies were evaluated. The arithmetic mean, the maximum and the summation. As can be seen in Figure 7 A) all three merging strategies have similar performances in terms of the c-index using variance thresholding and recursive feature elimination. The summation, which was used for the best-performing pipeline, shows a slightly higher median and a lower variance. The statistical analysis resulted in an F-statistic of 0.4251 with a corresponding p-value of 0.6592. This result indicates that there are no statistically significant differences among the three aggregation methods. It is relevant to mention that this subsection as well as the ones following are based on the results of different experiments, and not only one value is changing per test, e.g., different features are selected for each.

#### *4.1.3. Evaluation of Merging Points*

The comparison of the c-index using image-wise and metastases-wise in combination with variance thresholding and recursive feature elimination for feature selection resulted in a t-statistic of 0.0425 and a p-value of 0.9665. Hence, there is no statistically significant difference. This finding is further supported by Figure 7 B) which indicates similar performances.

#### *4.1.4. Evaluation of Segmentations*

The performance difference with regards to the c-index is not significant, as shown in Figure 7 C) and the difference might be caused by a single result. A t-statistic of $-0.1275$ and a p-value of 0.8997 show no statistically significant difference. It is worth pointing out that the best-performing pipeline used manually adjusted masks. Nevertheless, the performance using InteractiveNet's segmentation resulted in the second-best-performing pipeline, which was worse by a c-index of 0.005 ($0.7457 \pm 0.1281$ vs $0.7507 \pm 0.1373$). Taking the standard deviations into account, the difference between the methods is not statistically significant. In this context, it is important to consider that the manual adjustment was made to the result of InteractiveNet. For example, Figure 3 shows a tumor (2nd from the left) where the manual adjustment is identical with the output of InteractiveNet as no adjustments were necessary. Figure 3 B) displays the median difference between the two segmentations based on the absolute number of voxel differences in the metastasis. The adjustment required 711 voxel changes. In the case with the greatest dif-

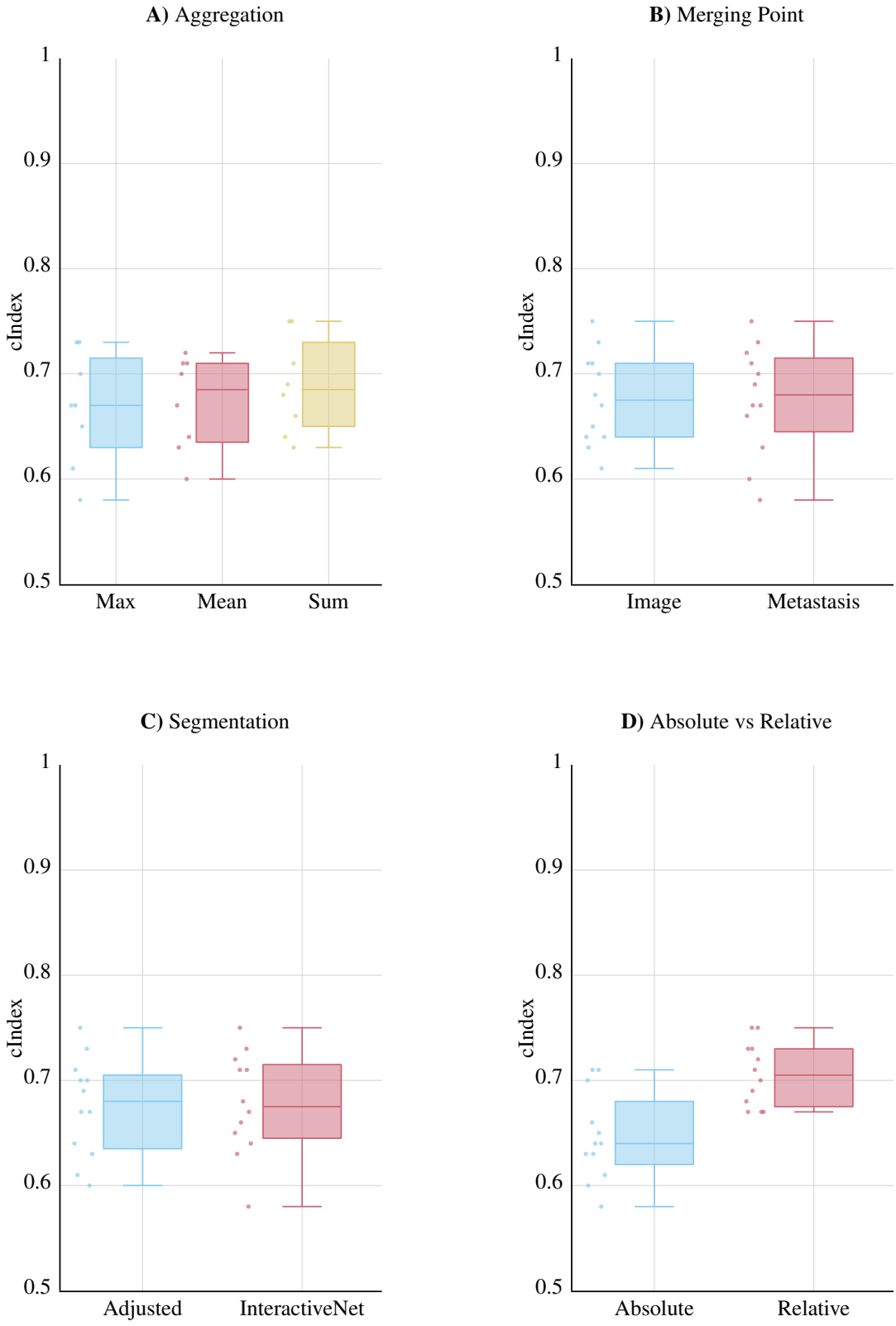



Figure 7: Boxplots for a variety of comparisons, utilizing the c-indices from nested cross-validation combined with variance thresholding and recursive feature elimination for feature selection. The underlying data points are presented as dots next to the boxplots.

Table 4: The selected features of the best-performing dataset of the respective feature selection methods. Features that appear more than once are highlighted in bold.

| Category | Var and RFECV | Var and repeated RFECV |
|---|---|---|
| Shape | **Mesh-Volume** | **Mesh-Volume** |
| | | Surface-Area |
| Firstorder | **Uniformity** | **Uniformity** |
| | Skewness | |
| glcm | Cluster-Prominence | Maximum-Probability |
| | Joint-Energy | |
| gldm | Small-Dependence-Low-Gray-Level-Emphasis | |
| glrlm | High-Gray-Level-Run-Emphasis | |
| glszm | Large-Area-High-Gray-Level-Emphasis | Small-Area-High-Gray-Level-Emphasis |
| | High-Gray-Level-Zone-Emphasis | |
| ngtdm | Complexity | |

ference, segmentation varied by 24,169 voxels, as illustrated in Figure 3 C). In some cases, manual adjustments resulted in potential oversegmentaions compared to InteractiveNet's segmentation, as seen in the first plot on the bottom row. Despite this outlier, the manually adjusted segmentations can still be considered better, especially for metastases outside of the lung where the contrast is significantly less.

#### 4.1.5. *Evaluation of Absolute and Relative Delta Features*

The mean c-index for absolute delta features was 0.663 with a standard deviation of 0.029, while the mean c-index for relative delta features was 0.704 with a standard deviation of 0.022. The boxplot is shown in Figure 7 D), highlighting a potentially better performance using relative delta features. The t-test confirmed this by showing a significant difference between the two methods, with a t-statistic of $-3.881$ and a p-value of 0.0008, indicating that relative delta features provides a statistically significantly better performance than absolute delta features.

### 4.2. *SHAP*

Figure 8 displays the patient-wise shap values of a single feature, the mesh-volume, over time. At least two subgroups can be seen. The first shows a negative shap score when an extremely high image-wise tumor growth rate is present. This can be interpreted that the survival time shortens if the tumor growth is large. A high growth rate is, in this case, a volume increase in the follow-up of over 200% compared to the baseline scan. Next, a "medium" growth rate correlates to a positive shap score in the second group. The model indicates that patients with up to 15% tumor growth may have a longer survival time. This finding aligns with RECIST guidelines, which define a similar subgroup with stable progression, characterized by less than 20% growth in tumor diameter. The third possible group is characterized by a tumor growth of more than 40%, but less than 200%. The shap values indicate that this does not change the survival time from the mean. All these subgroups can only be seen in the first years, as they align near 4000 days. The most probable explanation for this is the reduced number of patients used for training at the respective time points. While at 1000 days, half of the patient population remains, only two patients are still alive after 4000 days. It is important to note that volume is not the only feature that influences the decision making of the model. Other features such as skewness and the glszm large-area-high-gray-level-emphasis seem to influence the decision-making of the model as well, judging from Figure 9. In this context, it was found that an increase in large-area-high-gray-level-emphasis is related to a lower survival time, while a decrease leads to improved survival time. Taking into account the time difference and expressing it in simpler terms, a lower survival time is present if the metastasis in the follow-up scan is brighter than in the baseline scan.

### 4.3. *Evaluation of performance results on test set*

Lastly, the results of the final model are evaluated on the test, achieving a c-index of 0.70. The performance of the final model in the nested cross-validation was 0.75 with a standard deviation of 0.14. Even though a performance drop is present, it is within the expected deviation. Using the same feature set and pipeline as the final model, but changing a single variable in the dataset, resulted in a worse performance for each method tested (see Table 5). For example, changing the segmentation to InteractiveNet resulted in a c-index of 0.65.

Using the final model trained on manually adjusted segmentations and testing on InteractiveNet segmentations led to a c-index of 0.7619. This is the highest performance overall. Figure 10 indicates that the difference in performance, compared to the test using manually adjusted segmentations, originates from a vastly changed prediction for patient 6. To determine the influence of this change, the c-index was recalculated using the prediction of InteractiveNet for patient 6, but using the predictions with adjusted segmentations for

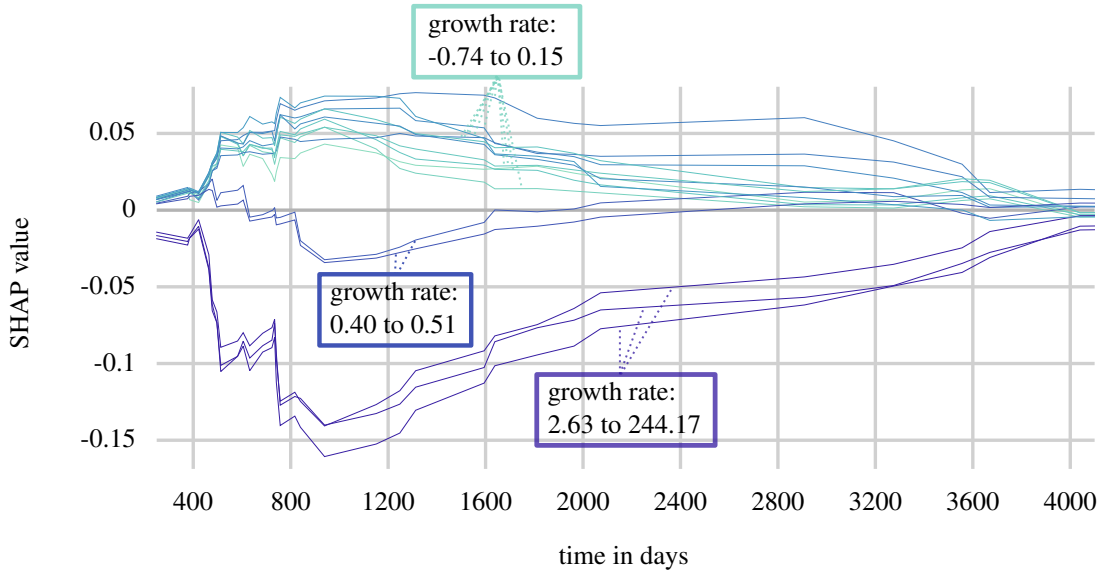



Figure 8: SHAP values for mesh volume, with each line representing a patient and color-coded according to metastasis growth rate. The color spectrum ranges from purple, indicating high growth rates, to various shades of blue and turquoise, representing low or negative growth rates.

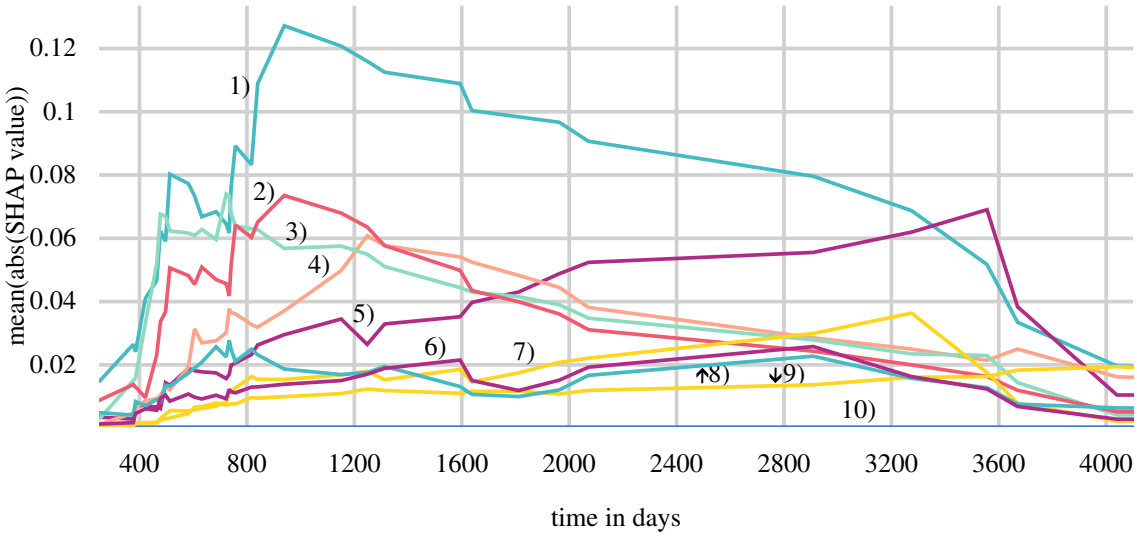



Figure 9: For every feature, the patient-wise mean of absolute shap values is shown. The features are 1) glszm Large-Area-High-Gray-Level-Emphasis, 2) Shape Mesh-Volume, 3) ngtdm Complexity, 4) Firstorder Skewness, 5) glcm Cluster-Prominence, 6) glszm High-Gray-Level-Zone-Emphasis, 7) gldm Small-Dependence-Low-Gray-Level-Emphasis, 8) glrlm High-Gray-Level-Run-Emphasis, 9) glcm Joint-Energy and 10) Firstorder Uniformity.

Table 5: Performance results with varying pipeline configurations.

| **Configuration** | **Variable Changed** | **c-index** (test set) |
|---|---|---|
| Adjusted, Sum, Relative, Image | None | **0.7024** |
| Adjusted, Sum, Relative, Metastasis | Metastasis | 0.6548 |
| Adjusted, Sum, Absolute, Image | Absolute | 0.6426 |
| InteractiveNet, Sum, Relative, Image | InteractiveNet | 0.6464 |

all other patients. This resulted in a c-index of 0.7619, confirming that the entire difference in performance was caused by this patient. Upon closer inspection of this patient, the difference has been tracked down to three slices of one metastasis in the lung. In those slices, InteractiveNet undersegments the metastasis, while the manual adjusted oversegments it by including a padding of not more than 3 millimeters of lung in every direction. While the change in volume is less than 1%, the texture features experience changes of up to 500%.

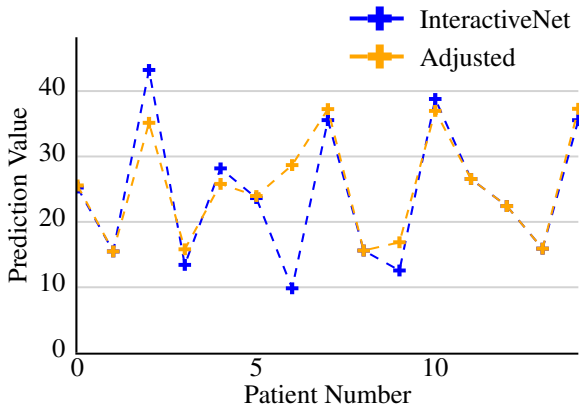



Figure 10: Prediction values for each patient in the test set using either InteractiveNet or manually adjusted segmentations.

## 5. Discussion

In the following section, the results will be critically discussed. Initially, the final model, its clinical applicability and its interpretability will be highlighted. Next, a comparison of segmentations, merging points, aggregation strategies, and both absolute and relative delta features will be presented. Finally, the limitations will be addressed.

### *5.1. The Final Model*

The final model shows a decent performance in terms of c-index and manages to identify various subgroups. The performance of the model is not perfect with a c-index of 1, but has a performance of 0.70. This might be due to the model itself, the dataset size and the metric used. In this dataset, several patients died within days of each other. However, the c-index assigns the same importance to differentiating these patients as compared to patients who are one year apart in survival time. This is important to consider, as the latter may be more relevant from a clinical point of view. In future studies, it might be useful to weight the patients according to their time difference for the calculation of the c-index or to exclude patients from the metric patients whose survival time is less than a certain time difference. Considering these points, other factors should be taken into account to evaluate model performance, such as the ability of the model to identify subgroups or the relation of the prediction to clinical observations, such as the RECIST guidelines. As both factors were found and a reasonable c-index was achieved, the model successfully demonstrated the feasibility of predicting therapy response in metastatic STS patients using delta radiomics.

With regards to the model itself, the performance difference of the test set compared to the nested cross-validation has to be discussed. Even though, nested cross-validation was used to reduce the overfitting due to model selection bias, a difference is notable. This could be due to the exclusion of the feature selection method inside the nested cross-validation. Another explanation can be found by taking a look at the patient characteristics in Table 1. As noted before, soft tissue sarcomas are a very diverse group of cancers. In this dataset, eleven different phenotypes were present. Nonetheless, the distribution of phenotypes seems to be rather similar in the training set used for the nested cross-validation and the test set. A bigger difference can be found in the locations of metastasis. In the training distribution, 14% of scans feature metastases in multiple sites, compared to 46% in the test distribution. As metastases in similar locations are more likely to have similar features, the gap between training and test performance might be caused by this fact. Adding to that, there is a vast difference in five year survival. While only 21% of patients are alive after five years, more than 36% are alive in the test set. Despite the challenges mentioned, the performance difference is not statistically significant.

#### *5.1.1. SHAP*

The existence of subgroups with regards to the volume was shown before. It is worth pointing out that it is not clear how many subgroups are present. Despite this, the characteristics of the presented subgroups have shown overlap with the RECIST guidelines. As

**The Final Model**

*Dataset*

| | |
|---|---|
| **Segmentation** | Adjusted |
| **Aggregation** | Summation |
| **Abs or Rel** | Relative |
| **Merged at** | Image |
| **Feature Selection** | Var and RFECV |

*Pipeline*

| | |
|---|---|
| **Outlier Detection** | Interquartile range |
| **Scaler** | Quantile transformer |
| **Survival Model** | Random survival forest |

*c-index*

| | |
|---|---|
| **nested cross-validation** | $0.75 \pm 0.14$ |
| **test set** | 0.7024 |

Figure 11: The score sheet of the final model

RECIST is solely based on the diameter and fails to correctly predict therapy response, one might argue that the presented final model should fail as well. However, the volume is used instead of the diameter and more importantly, the volume is used in combination with nine other image features. In combination, this feature set might provide better therapy response prediction, as indicated by its good performance. One of the features is the large-area-high-gray-level-emphasis through which it was found that an increase in brightness of the metastasis over time leads to a lower survival time. An increase in brightness might be related to a higher cellularity or necrosis. As the shap values are available for each patient prediction, they could also be used in a clinical setting to enhance the interpretability of the model.

### *5.2. Comparison of Segmentations*

In the automatic machine learning, the two segmentations show no statistically significant difference in terms of the performance measured by the c-index. On the first look, this suggests that the segmentations are interchangeable. However, different feature sets were selected. Using the feature sets of adjusted segmentations, but training and testing with InteractiveNets segmentations resulted in a worse performance. This indicates that the selected features are linked to the segmentations and therefore the segmentations are not interchangeable. Adding up on that, the prediction of the final model drastically changed when a few millimeters of lung were oversegmented. This suggests that voxel perfect segmentations might be needed in clinical practice in order to get reproducible and stable results. However, it is worth pointing out that the model seems to be able to handle differences in segmentations to some extent, as only one outlier was found.

### *5.3. Comparison of Different Merging Strategies*

While the three merging strategies have no significant differences, the interpretation of the features changes drastically between them. This is mainly due to the fact that the different merging strategies incorporate different information of the metastases. Taking the mesh-volume and metastasis-wise merging as an example, the mean would correspond to the average metastasis growth, hence indirectly incorporating information from all metastases. In comparison. the maximum value refers to the largest change in a specific radiomic feature. Each feature value corresponds to a single metastasis. Since multiple features are extracted, different metastases might show the largest change for different features. For instance, the maximum volume change might come from one metastasis, while the maximum uniformity change might come from another. This also implies that all metastases must be segmented to accurately reflect the changes. The summation, on the other hand, includes information about all metastases directly. It is worth pointing out that using the summation, a metastasis growth of, e.g., $100\,mm$ does not necessarily correspond to the total growth as not all metastases have been segmented in order to follow RECIST guidelines. This might lead to worse results because only the information of a few metastases are incorporated. In comparison to that, one could argue that the mean of the segmentations is the mean of all metastases. However, this might only be true for texture features, but not for shape features such as mesh volume, as the biggest metastases are segmented. This might explain why summation was used for the final model and not the mean, as expected. It is certain that more tests are needed to accurately differentiate the merging strategies. For example, one might consider repeating the experiments using segmentations from all metastases.

### *5.4. Comparison of Different Merging Points*

With the hypothesis that the metastasis-wise merging was the superior choice, it is surprising to see image-wise merging being used for the final model. Adding up on that, it is worth mentioning that no significant difference between the two over all tested combinations was found. Despite the fact that this result is not directly associated with a clinical interpretation, it is very important for the creation of feature studies. This is because the image-wise merging does not require matching tumor labels anymore, which simplifies the annotation problem. Furthermore, it resolves the problem of how a physically missing tumor or a new tumor in the follow-up can be imputed. Here, a feature space of zeros was chosen, which might not reflect a fully treated,

hence missing, tumor at all. This effect might explain why metastasis-wise merging does not perform better than image-wise merging, even though metastasis-wise merging is much more logical from a clinical point of view as results could be backtracked to a specific metastasis rather than the entire image. It is certain that more tests are needed to accurately differentiate the merging strategies.

### 5.5. Comparison of Absolute and Relative Delta Features

While the absolute difference is the method of choice for all the authors presented in section 2, it does not outperform the relative delta features. On the contrary, the relative delta features perform statistically better. This speaks for the interpretation that relative tumor growth is more important than, e.g., absolute tumor growth. This fact aligns with the findings of RECIST, which state relative differences rather than absolute ones. However, RECIST notes that a tumor has to be a certain absolute size before it can be used for interpretation. Following this logic, it might be useful to merge absolute and relative features into one feature space instead of treating them separately. Additional improvements could be made by normalizing the delta features with the time in between the scans. This might be of importance as a tumor growth rate of $\frac{10\%}{month}$ over three months is significantly different compared to a growth of four months.

### 5.6. Limitations

In the following, the limitations of the work are presented. For this, the used metrics, automatic machine learning method and methodological choice are discussed.

#### 5.6.1. Limitations of the Metrics

Even though the c-index is arguably the most common metric for survival prediction, it has several drawbacks. First, only patients for whom the event occurred are included. This reduces the number of patients used for model evaluation substantially. Because the c-index is used for selecting the best pipeline in the nested cross-validation, the relevance of patients who did not die is neglected. Second, the concordance index is a patient based metric of ordering. This might not be the most useful, because clinicians might be more interested in subgroups. An example by Hartman et al. (2023) suggests considering a group of ten-year-olds and a group of ninety-year-olds. Without doubt, it can be said that the survival time of the ten year old's is much higher and they thus have a lower risk of death. Given a large enough dataset, one should be easily able to train a model that perfectly predicts a higher survival time for all ten-year-olds. As the model is able to make this prediction perfectly, one might guess that the concordance index is 1. However, for a perfect c-index, the ten-year-old's need to be perfectly sorted in between themselves. The same has to be true for the ninety-year-old's. This is undeniably a much harder task and might result in a lower c-index. In the extreme case of balanced groups and age being the only feature, this would result in a c-index of 0.75. Despite the fact that the model makes a perfect prediction from a clinician's point of view that one patient population always lives longer than another, this is not reflected in a c-index of 1. In the case of an imbalanced dataset, i.e., 80% are 10-year-olds, the c-index would only be 0.66 despite a perfect discrimination of the underlying subgroups. Coming back to the task at hand, it is important to consider potential subgroups and not only look at the c-index. While other metrics have been proposed to overcome the problem of censoring in the c-index, such as the c-index based on the inverse probability of censored weights, the much more severe problem of the subgroups remains. Furthermore, the other metrics require that the extreme survival times of the test set are in between the extreme survival times of the training set. Especially in cross-validation settings, random splitting would not be possible anymore under these circumstances.

#### 5.6.2. Limitations of Automatic machine learning

One limitation here is the use of feature selection methods. Due to rather long selection times, they were not included in the nested cross-validation at the cost of a selection bias. Furthermore, the repetition of the recursive feature elimination with cross-validation should not have led to better results compared to a single run for any dataset. The fact that it did speaks for the randomness of the feature selection process. This is despite the best efforts to reduce these effects by using seeds for the generation of pseudo-random numbers.

#### 5.6.3. Limitations of the Methodology

In this study, 58 patients were included, 15 of whom were in the test set. Given the size, a change in the prediction of a single patient has led to a performance difference of 6%. As it is difficult to analyze robustness under these circumstances, a larger patient population is needed. Additionally, patients from multiple centers should be included to analyze effects such as feature robustness and scanner inter variability.

Another constraint is the segmentation of only the five biggest tumors after the RECIST guidelines. In future studies, it might be reasonable to test the performance difference when all metastases are segmented. However, this would increase the segmentation time considerably. Finally, this study deliberately chose to predict the therapy response directly from overall survival rather than converting it to a classification task. While this method is certainly feasible, judging from the results presented before, it has some drawbacks as well as advantages. One point to consider here is the model

output. On the one hand, the continuous prediction of the survival model holds more information than the prediction of the classification task. On the other hand, the interpretation of a continuous label, especially one of arbitrary scale, might be more difficult. In the future, it might be useful to compare the presented results to a classification based, e.g., on three-year overall survival.

## 6. Conclusions

Summing it all up, multiple conclusions can be drawn. 1) The presented method shows the feasibility of predicting therapy response exclusively on metastatic STS with a c-index of 0.70. 2) Relative delta features show statistically significant better performance compared to absolute delta features. 3) Image-wise merging is a suitable substitute for metastasis-wise merging and simplifies the annotation process as no matching labels are required.

## Acknowledgments

First, I would like to thank D.S. and M.S. for their continued guidance and support. Further thanks are due to all the clinicians and staff involved in the production of this dataset. A special thanks goes to M.V. for the segmentation of more than one-hundred CT scans with over three-hundred individual sarcoma segmentations.

## Acronyms

**AUC** area under the curve.

**c-index** harrell's concordance index.

**CT** Computed Tomography.

**ESMO** European Society for Medical Oncology.

**glcm** gray-level co-occurrence matrix.

**gldm** gray-level dependence matrix.

**glrlm** gray-level run-length matrix.

**glszm** gray-level size-zone matrix.

**MRI** Magnetic Resonance Imaging.

**ngtdm** neighboring-gray-tone difference matrix.

**STS** soft tissue sarcoma.

## References

Coindre, J.M., Terrier, P., Guillou, L., Le Doussal, V., Collin, F., Ranchère, D., Sastre, X., Vilain, M.O., Bonichon, F., N'Guyen Bui, B., 2001. Predictive value of grade for metastasis development in the main histologic types of adult soft tissue sarcomas. Cancer 91, 1914–1926. doi:10.1002/1097-0142(20010515)91:10<1914::AID-CNCR1214>3.0.CO;2-3.

Cousin, F., Louis, T., Dheur, S., Aboubakar, F., Ghaye, B., Occhipinti, M., Vos, W., Bottari, F., Paulus, A., Sibille, A., Vaillant, F., Duysinx, B., Guiot, J., Hustinx, R., 2023. Radiomics and Delta-Radiomics Signatures to Predict Response and Survival in Patients with Non-Small-Cell Lung Cancer Treated with Immune Checkpoint Inhibitors. Cancers 15. doi:10.3390/cancers15071968.

Crombé, A., Périer, C., Kind, M., De Senneville, B.D., Le Loarer, F., Italiano, A., Buy, X., Saut, O., 2019. T2-based MRI Delta-radiomics improve response prediction in soft-tissue sarcomas treated by neoadjuvant chemotherapy. Journal of Magnetic Resonance Imaging 50, 497–510. doi:10.1002/jmri.26589.

Crombé, A., Spinnato, P., Italiano, A., Brisse, H.J., Feydy, A., Fadli, D., Kind, M., 2023. Radiomics and artificial intelligence for soft-tissue sarcomas: Current status and perspectives. Diagnostic and Interventional Imaging 104, 567–583. doi:10.1016/j.diii.2023.09.005.

Eisenhauer, E.A., Therasse, P., Bogaerts, J., Schwartz, L.H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., Rubinstein, L., Shankar, L., Dodd, L., Kaplan, R., Lacombe, D., Verweij, J., 2009. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). European Journal of Cancer 45, 228–247. doi:10.1016/j.ejca.2008.10.026.

Fields, B.K.K., Demirjian, N.L., Cen, S.Y., Varghese, B.A., Hwang, D.H., Lei, X., Desai, B., Duddalwar, V., Matcuk, G.R., 2023. Predicting Soft Tissue Sarcoma Response to Neoadjuvant Chemotherapy Using an MRI-Based Delta-Radiomics Approach. Molecular Imaging and Biology 25, 776–787. doi:10.1007/s11307-023-01803-y.

Gao, Y., Kalbasi, A., Hsu, W., Ruan, D., Fu, J., Shao, J., Cao, M., Wang, C., Eilber, F.C., Bernthal, N., Bukata, S., Dry, S.M., Nelson, S.D., Kamrava, M., Lewis, J., Low, D.A., Steinberg, M., Hu, P., Yang, Y., 2020. Treatment effect prediction for sarcoma patients treated with preoperative radiotherapy using radiomics features from longitudinal diffusion-weighted MRIs. Physics in Medicine & Biology 65, 175006. doi:10.1088/1361-6560/ab9e58.

Gronchi, A., Miah, A.B., Dei Tos, A.P., Abecassis, N., Bajpai, J., Bauer, S., Biagini, R., Bielack, S., Blay, J.Y., Bolle, S., Bonvalot, S., Boukovinas, I., Bovee, J.V.M.G., Boye, K., Brennan, B., Brodowicz, T., Buonadonna, A., De Álava, E., Del Muro, X.G., Dufresne, A., Eriksson, M., Fagioli, F., Fedenko, A., Ferraresi, V., Ferrari, A., Frezza, A.M., Gasperoni, S., Gelderblom, H., Gouin, F., Grignani, G., Haas, R., Hassan, A.B., Hecker-Nolting, S., Hindi, N., Hohenberger, P., Joensuu, H., Jones, R.L., Jungels, C., Jutte, P., Kager, L., Kasper, B., Kawai, A., Kopeckova, K., Krákorová, D.A., Le Cesne, A., Le Grange, F., Legius, E., Leithner, A., Lopez-Pousa, A., Martin-Broto, J., Merimsky, O., Messiou, C., Mir, O., Montemurro, M., Morland, B., Morosi, C., Palmerini, E., Pantaleo, M.A., Piana, R., Piperno-Neumann, S., Reichardt, P., Rutkowski, P., Safwat, A.A., Sangalli, C., Sbaraglia, M., Scheipl, S., Schöffski, P., Sleijfer, S., Strauss, D., Strauss, S., Sundby Hall, K., Trama, A., Unk, M., van de Sande, M.a.J., van der Graaf, W.T.A., van Houdt, W.J., Frebourg, T., Casali, P.G., Stacchiotti, S., ESMO Guidelines Committee, EURACAN and GENTURIS. Electronic address: clinicalguidelines@esmo.org, 2021. Soft tissue and visceral sarcomas: ESMO-EURACAN-GENTURIS Clinical Practice Guidelines for diagnosis, treatment and follow-up. Annals of Oncology: Official Journal of the European Society for Medical Oncology 32, 1348–1365. doi:10.1016/j.annonc.2021.07.006.

Hartman, N., Kim, S., He, K., Kalbfleisch, J.D., 2023. Pitfalls of the concordance index for survival outcomes. Statistics in Medicine 42, 2179–2190. doi:10.1002/sim.9717.

Hoaglin, D., Iglewicz, B., 1993. Volume 16: How to Detect and Handle Outliers.

Krzyziński, M., Spytek, M., Baniecki, H., Biecek, P., 2023. SurvSHAP(t): Time-dependent explanations of machine learning survival models. Knowledge-Based Systems 262, 110234. doi:10.1016/j.knosys.2022.110234.

Lochner, J., Menge, F., Vassos, N., Hohenberger, P., Kasper, B., 2020. Prognosis of Patients with Metastatic Soft Tissue Sarcoma: Advances in Recent Years. Oncology Research and Treatment 43, 613–619. doi:10.1159/000509519.

Meyer, M., Seetharam, M., 2019. First-Line Therapy for Metastatic Soft Tissue Sarcoma. Current Treatment Options in Oncology 20, 6. doi:10.1007/s11864-019-0606-9.

Miao, L., Cao, Y., Zuo, L., Zhang, H., Guo, C., Yang, Z., Shi, Z., Jiang, J., Wang, S., Li, Y., Wang, Y., Xie, L., Li, M., Lu, N., 2023. Predicting pathological complete response of neoadjuvant radiotherapy and targeted therapy for soft tissue sarcoma by whole-tumor texture analysis of multisequence MRI imaging. European Radiology 33, 3984–3994. doi:10.1007/s00330-022-09362-6.

Nardone, V., Reginelli, A., Guida, C., Belfiore, M.P., Biondi, M., Mormile, M., Banci Buonamici, F., Di Giorgio, E., Spadafora, M., Tini, P., Grassi, R., Pirtoli, L., Correale, P., Cappabianca, S., Grassi, R., 2020. Delta-radiomics increases multicentre reproducibility: A phantom study. Medical Oncology 37, 38. doi:10.1007/s12032-020-01359-9.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Peeken, J.C., Asadpour, R., Specht, K., Chen, E.Y., Klymenko, O., Akinkuoroye, V., Hippe, D.S., Spraker, M.B., Schaub, S.K., Dapper, H., Knebel, C., Mayr, N.A., Gersing, A.S., Woodruff, H.C., Lambin, P., Nyflot, M.J., Combs, S.E., 2021. MRI-based delta-radiomics predicts pathologic complete response in high-grade soft-tissue sarcoma patients treated with neoadjuvant therapy. Radiotherapy and Oncology 164, 73–82. doi:10.1016/j.radonc.2021.08.023.

Pölsterl, S., 2020. Scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. Journal of Machine Learning Research 21, 1–6.

Qu, H., Zhai, H., Zhang, S., Chen, W., Zhong, H., Cui, X., 2023. Dynamic radiomics for predicting the efficacy of antiangiogenic therapy in colorectal liver metastases. Frontiers in Oncology 13, 992096. doi:10.3389/fonc.2023.992096.

Spaanderman, D.J., Starmans, M.P.A., van Erp, G.C.M., Hanff, D.F., Sluijter, J.H., Schut, A.R.W., van Leenders, G.J.L.H., Verhoef, C., Grunhagen, D.J., Niessen, W.J., Visser, J.J., Klein, S., 2024. Minimally Interactive Segmentation of Soft-Tissue Tumors on CT and MRI using Deep Learning. doi:10.48550/arXiv.2402.07746, arXiv:2402.07746.

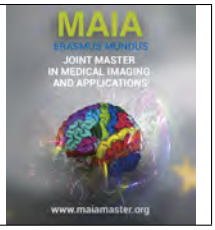
Stacchiotti, S., Collini, P., Messina, A., Morosi, C., Barisella, M., Bertulli, R., Piovesan, C., Dileo, P., Torri, V., Gronchi, A., Casali, P.G., 2009. High-Grade Soft-Tissue Sarcomas: Tumor Response Assessment—Pilot Study to Assess the Correlation between Radiologic and Pathologic Response by Using RECIST and Choi Criteria. Radiology 251, 447–456. doi:10.1148/radiol.2512081403.

Stiller, C.A., Botta, L., Brewster, D.H., Ho, V.K.Y., Frezza, A.M., Whelan, J., Casali, P.G., Trama, A., Gatta, G., 2018. Survival of adults with cancers of bone or soft tissue in Europe—Report from the EUROCARE-5 study. Cancer Epidemiology 56, 146–153. doi:10.1016/j.canep.2018.08.010.

van Griethuysen, J.J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G., Fillion-Robin, J.C., Pieper, S., Aerts, H.J., 2017. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer research 77, e104–e107. doi:10.1158/0008-5472.CAN-17-0339.

Medical Imaging and Applications

Master Thesis, June 2024




# Deep Learning-Based Detection of Homologous Recombination Deficiency (HRD) in Ovarian Cancer Whole Slide Histopathology Images

Md Imran Hossain[a], Patrick Bard[a], Valentin Derangère[b,c,d], Caroline Truntzer[b,c], Michel Paindavoine[a], Manon Ansart[a]

[a] *LEAD-CNRS, Université de Bourgogne, 21000 Dijon, France*
[b] *Centre Georges-François Leclerc, 21000 Dijon, France,*
[c] *INSERM U1231 équipe Tirecs, 21000 Dijon, France,*
[d] *UFR des Sciences de Santé, Université de Bourgogne,2100 Dijon, France*

**Abstract**

Breast and ovarian cancers are among the most common cancer diseases affecting women worldwide, which creates a significant global health challenge. Early diagnosis and the application of precision medicine are pivotal in addressing these challenges, as they enable personalized treatment plans and effective therapeutic strategies. Identifying predictive biomarkers, such as Homologous Recombination Deficiency (HRD), is crucial since HRD-positive tumors are highly responsive to certain therapies. While several studies have demonstrated promising results for HRD prediction in breast cancer using deep learning, its application to ovarian cancer remains underexplored. In this study, we used a deep learning-based pipeline for detecting HRD from whole-slide digital histopathology images (WSIs) of ovarian cancer. We also applied the pipeline to breast cancer WSIs considering a standard validation, as several studies demonstrated considerable performance. We employed various techniques, including fully supervised, weakly supervised, and self-supervised with transfer learning, across public and private datasets. Our comparative analysis of these methods revealed a mean Area Under the Curve (AUC) of 0.66 for the TCGA-OV dataset and an AUC of 0.55 for the DIJON-OV dataset. Conversely, the mean AUC of 0.77 for the TCGA-BRCA dataset highlights the robustness of the pipeline for HRD prediction in breast cancer. This research underscores the capability of deep learning-based approaches in predicting HRD in breast cancer WSIs, while also indicating the need for further studies to improve their efficacy in ovarian cancer.

*Keywords:* Homologous Recombination Deficiency, Digital Pathology, Deep Learning, Ovarian Cancer.

## 1. Introduction

Cancer is a life-threatening disease and its severity is increasing day by day (Ohlén and Holm, 2006). Every year, millions of individuals suffer from various forms of cancer (Sung et al., 2021), and a significant portion succumb due to several reasons including late diagnosis, inadequate treatment facilities, and the aggressive nature of their disease. Just in 2020, there were an estimated 19.3 million new cancer cases and nearly 10 million cancer-related deaths (Sung et al., 2021). Both men and women were significantly affected by this devastating disease. Among women, ovarian cancer is one of the prevalent types, with around 313,000 new cases reported annually (Huang et al., 2022; Sung et al., 2021). These statistics highlight the critical global health challenge presented by cancer, emphasizing the urgent need for better diagnostic methods, more effective treatments, and increased awareness to fight this pervasive disease. However, early diagnosis and precision medicine can play a significant role in mitigating these challenges (Fitzgerald et al., 2022; Yang et al., 2023). Women suffering from ovarian cancer can greatly benefit from personalized treatment plans and effective therapeutic strategies. (Chan et al., 2017; Fernandez-Garza et al., 2021). Therefore, identifying predictive biomarkers, such as Homologous Recombination Deficiency (HRD), is essential to guide the de-

velopment of effective and personalized treatment plans for ovarian cancer patients (Lenz et al., 2023; Ngoi and Tan, 2021; Shi et al., 2021).

Homologous Recombination (HR) is a cellular process in which two similar or identical DNA molecules exchange genetic information through the pairing and exchange of nucleotide sequences (Liu and Konstantinopoulos, 2017; Qi et al., 2015). HR is a crucial biological process as it repairs the double-strand breaks caused by various factors including radiation, chemical, and normal cellular processes (Gelot et al., 2016; Li and Heyer, 2008). The deficiency of the HR process, also known as HRD, results in improper or error-prone repairs of Double-Strand Breaks. This deficiency is a consequence of BRCA1/2 genetic mutations, associated with ovarian cancer (Lazard et al., 2022; Miller et al., 2020). HRD induced by BRCA1 and BRCA2 is highly sensitive and responsive to specific therapies, such as platinum-based chemotherapy (Tutt Andrew N.J. et al., 2021) and polyADP-ribose polymerase (PARP) inhibitors (Tutt et al., 2018). Hence, the HR status (deficiency or proficiency) can serve as a predictive biomarker that significantly aids in treatment planning and therapeutic decision-making for ovarian cancer (Miller et al., 2020) and breast cancer (Chopra et al., 2020). Several methods including genomic instability profiling, mutational signatures, or integrating structural and mutational signatures have been used to detect HR status (Abkevich et al., 2012; Birkbak et al., 2012; Davies et al., 2017; Popova et al., 2012). However, these approaches require advanced laboratory infrastructures and high financial resources. Only a few laboratories have the capability to perform these experiments across the world. To overcome these challenges, we hypothesize that HR status can be predicted from the Hematoxylin and Eosin (H&E) stained tissue slides, commonly used in clinical pathology to assess tissue morphology and detect various abnormalities (Lahiani et al., 2018), with the help of deep learning.

Deep Learning has revolutionized biomedical image analysis in particular digital pathology (Deng et al., 2020). Most techniques in this field have been focused on computer-aided diagnosis, where the goal is to partially automate the human interpretation of slides to assist pathologists in their diagnostic tasks, such as identifying metastatic axillary lymph nodes (Campanella et al., 2019; Ehteshami Bejnordi et al., 2017) or detecting mitoses (Veta et al., 2015). Deep Learning has shown effectiveness not just in automating manual inspection but also in predicting patient factors like patient outcome and biological features, for example, gene mutations (Coudray et al., 2018), expression levels (Schmauch et al., 2020), and genetic signatures (Diao et al., 2021). However, Deep Learning also has some drawbacks due to its black-box nature. The lack of biological interpretation and validation resists the trustworthiness of making clinical decisions.

In this study, we aimed to predict the HR status from WSIs of ovarian cancer using deep learning techniques. We utilized previously developed models by other intern students and incorporated a new pipeline from Filiot et al. (2023) that has proven effective for classification tasks in pathology. The main focus of our study was on HRD detection in ovarian cancer. However, we also included breast cancer data for validation purposes to confirm the effectiveness of the pipeline, since HRD detection in breast cancer is well-established, whereas it is still less certain in ovarian cancer. The study presents the following activities:

1. *Fully supervised method:*
    - We used a pre-developed pipeline by other intern students that employs Convolutional Neural Networks (CNNs) models, specifically CNN, ResNet34, and ResNet50, to predict HR status from breast and ovarian cancer WSIs. The evaluation of this method was conducted on both public and private datasets, ensuring a comprehensive assessment of its performance.
2. *Combination of self-supervised and weakly supervised method:*
    - We integrated a new pipeline developed by Filiot et al. (2023) that combines self-supervised and weakly supervised methods using transfer learning techniques.
    - A Vision Transformer (ViT) based pre-trained model trained with self-supervised learning methods, such as Image BERT Pre-Training with Online Tokenizer (iBOT) and Masked Image Modeling (MIM) , on large datasets was utilized to extract features from breast and ovarian cancer WSIs.
    - These extracted features were then used in weakly supervised methods, such as Multiple Instance Learning (MIL) models, to predict HR status.
3. *Comparative study:*
    - A comparative study highlighted the effectiveness of fully supervised, weakly supervised, and self-supervised method with transfer learning techniques in predicting HR status. This analysis provided valuable insights into the strengths and limitations of each method, contributing to a better understanding of their applicability in HRD prediction in ovarian cancer.

By integrating these methodologies, our study demonstrates an approach for predicting HR status from ovarian cancer WSIs, highlighting the potential of deep learning techniques in computational pathology. Validation with breast cancer data confirmed the reliability of the pipelines used.

## 2. State of the art

Recent advancements in deep learning and computational pathology have demonstrated significant potential for improving cancer subtype classification and tumor histopathology evaluation. WSIs, digitized at high resolutions with dimensions ranging from 10,000 to over 150,000 pixels, are too large for current graphics processing units (GPUs) to handle or process at their original size. As a result, WSIs are often divided into smaller tiles for analysis. Traditional fully supervised learning approaches in computational pathology often require extensive labeled data, which is impractical, time-consuming, and expensive (Marini et al., 2021). Pathologists must meticulously annotate each patch to distinguish between different tissues or classify various cancer subtypes. While this level of detail is crucial for training accurate models, it poses significant challenges in terms of time and cost. To address these issues, recent studies have proposed several weakly supervised learning and self-supervised learning methods. These approaches reduce the need for extensive annotations and are emerging as viable solutions to the limitations of fully supervised methods.

### *2.1. Fully supervised method*

The fully supervised approach in deep learning requires pathologists to provide manual tile-level annotations on a WSI, which is essential for training the models. Despite this meticulous, costly, and time-consuming task, fully supervised models remain one of the most effective and commonly employed strategies in state-of-the-art computational pathology for cancer subtype classification due to their high performance levels. Although this method has been effectively used for tasks such as detecting mitotic figures in breast cancer (Veta et al., 2015), identifying lymph node metastases (Campanella et al., 2019), and classifying various types of lung cancer (Yang et al., 2021), its application to predicting HR status in breast and ovarian cancers has been limited, highlighting a potential area for further research.

Cireşan et al. (2013) pioneered deep learning in computational pathology by applying deep neural networks to detect mitosis in breast cancer histology images, achieving an F1-score of 0.78 on the MITOS-ATYPIA 2014 dataset through data augmentation, patch extraction, and a sliding window approach. Later, Litjens et al. (2016) used a fully-supervised CNN to detect lymph node metastases in the CAMELYON16 dataset, achieving an Area Under the Curve (AUC) of 0.96 with preprocessing techniques such as color normalization and patch extraction. Janowczyk and Madabhushi (2016) used a fully supervised method with the AlexNet model (Krizhevsky et al., 2012) for mitosis detection, invasive ductal carcinoma detection, and lymphoma classification task and obtained F1-score of 0.53 across 550 mitotic events, F1-score of 0.76 on 50k testing patches, and an accuracy of 0.97 across 374 images respectively. Liu et al. (2017) utilized a CNN model for detecting breast cancer metastases in gigapixel pathology images. They achieved a tumor detection rate of 92.4% with 8 false positives per image and image-level AUC scores above 97% on the Camelyon16 dataset and an independent set of 110 slides. These studies collectively highlight the significant advancements in histopathology image analysis through fully supervised deep-learning methods.

### *2.2. Weakly supervised method*

While fully supervised approaches have demonstrated significant performance in computational pathology, they have notable limitations. This has led to the exploration of more robust techniques, such as weakly supervised methods. The weakly supervised method is a machine learning technique where the training data is only partially labeled. Instead of local labels for every instance (e.g., pixel-level or tile-level annotations), the annotations are provided at a global level (e.g., image-level or slide-level labels). This method leverages these less detailed labels to train models, making it a practical and cost-effective solution in scenarios where obtaining detailed annotations is challenging or expensive.

In computational pathology, weakly supervised methods more accurately reflect real-world scenarios, where a pathologist provides a single diagnosis per slide rather than detailed annotations for each tile. The slide-based or global levels are inherently noisy, as only a small region within a slide may be representative of the label. To address this issue, Multiple Instance Learning (MIL) has emerged as the state-of-the-art among weakly supervised algorithms.

#### *2.2.1. Multiple instance learning*

MIL effectively manages the noisy nature of slide-level annotations, enhancing the robustness and accuracy of the models. Within MIL frameworks, two different strategies for aggregating instance-level features into a bag-level representation are studied. These strategies aim to capture the key characteristics of cancer tissue samples and differentiate between various cancer types and normal cells.

- **Instance-level aggregation:** This approach involves building an instance-level classifier that assigns scores to each instance or tile within a slide. These scores are then aggregated using MIL pooling methods, such as max-pooling or mean-pooling. The pooling operation summarizes the information from each instance, capturing essential features associated with different cancer types. By aggregating these scores, the model effectively captures the distinct characteristics of the cancer tissue.

- **Embedding-level aggregation:** In this approach, each instance or tile is first mapped to a low-dimensional embedding. MIL pooling is then applied to these embeddings to obtain a single bag-level representation that is independent of the number of instances or tiles in each bag. This method allows the model to focus on the most relevant features within the embeddings, enhancing its ability to differentiate between cancerous and normal cells regardless of the variability in the number of patches.

#### 2.2.2. *Applications of MIL in computational pathology*

MIL has demonstrated significant potential in various applications within computational pathology, enhancing the accuracy and robustness of models for tasks such as cancer subtype classification and tumor grading (Anaya et al., 2024). Notably, the inclusion of the attention mechanism made the MIL method more strong and effective for histopathology classification using the slide-based labels. Ilse et al. (2018) introduced the noble attention-based MIL for breast and colon cancer classification tasks using WSIs and obtained 0.74 and 0.90 mean AUC respectively. Later, Courtiol et al. (2020) introduces Chowder MIL, a weakly supervised learning-based model, for disease localization in histopathology using only global labels of WSIs. This method contrasts with attention-based MIL by utilizing a combination of top-instance learning and negative evidence to effectively identify both the presence and absence of disease characteristics. Chowder MIL demonstrated its effectiveness with an impressive AUC of 0.87 on the Camelyon-16 challenge, efficiently identifying cancerous regions without detailed local annotations. Lu et al. (2021) proposed a data-efficient model known as Constrained-Attention Multiple-Instance Learning (CLAM) for computational pathology. This model introduced clustering features among relevant instances and is effective for both the binary and multi-class classification tasks. The CLAM model demonstrated superior performance over standard weakly supervised classification algorithms in several diagnostic tasks. It achieved a mean test AUC of 0.991 in subtyping renal cell carcinoma, 0.956 in subtyping non-small-cell lung cancer, and 0.953 in detecting lymph node metastasis. These results highlight the effectiveness of CLAM in accurately classifying complex digital pathology data across different cancer types. Shao et al. (2021) proposed the first transformer-based MIL model, known as Trans-MIL, which introduced the self-attention mechanism to include correlation among instances. This approach significantly outperformed traditional MIL-based models, achieving AUCs of up to 93.09% on the CAMELYON16 dataset and between 96.03% and 98.82% on the TCGA-NSCLC and TCGA-RCC datasets respectively.

#### 2.2.3. *MIL for HRD detection*

MIL methods can also be used for HRD detection from breast and ovarian cancer WSIs. Valieris et al. (2020) implemented a deep learning framework using CNN for feature extraction and MIL with a recurrent neural network for feature aggregation and classification from histopathological slides. This approach, aimed at detecting HRD in breast cancer, achieved an AUC of 0.80 on the TCGA dataset and an AUC of 0.70 on the independent dataset for validation. Nero et al. (2022) used CLAM for identifying HRD from ovarian cancer and obtained an AUC of 0.7 on the training set (464 slides) and 0.55 on the testing set (464 images). Bergstrom et al. (2023) developed the DeepHRD model, which uses a weakly supervised CNN and MIL to predict HRD from digital H&E slides. The model achieved an AUC of 0.81 on the TCGA breast cancer dataset and 0.76 on independent validation cohorts. For ovarian cancer, using transfer learning from the breast cancer data, the model showed significant predictive power by differentiating median survival times: HR-deficient (HRD) patients had a median survival of 4.6 years, while HR-proficient (HRP) patients had 3.2 years.

### 2.3. *Self-supervised method with transfer learning*

In recent years, self-supervised learning algorithms have gained prominence in computer vision, offering a powerful alternative to traditional supervised learning, which relies heavily on large annotated datasets. This reliance is particularly challenging in computational pathology due to the complexity of medical images and the high cost of expert annotations. Self-supervised learning generates labels from the data itself, enabling models to learn useful representations without extensive manual annotation. This reduces dependency on large labeled datasets and enhances model generalization from unstructured histopathological data.

In computational pathology, fully supervised methods require extensive annotated data and expert knowledge, which results in specialized datasets that are limited in size and often fail to generalize effectively. Weakly supervised methods reduce the annotation burden by using slide-level labels but typically underperform compared to fully supervised methods and struggle with accurately localizing disease regions within images, limiting their clinical applicability.

#### 2.3.1. *Advances in self-supervised learning*

Self-supervised learning addresses the mentioned issues by pre-training models using the dataset's own images and constructing tokenized dictionaries for unsupervised learning. For example, in natural language processing, tokenization involves breaking down text into smaller units called tokens, which facilitates efficient lookup and analysis. However, in the domain of computer vision, constructing these dictionaries poses

Table 1: Summary of studies on HRD detection using histopathology images in breast and ovarian cancer.

| Paper | Training Strategy | Dataset | Subtypes | Results | Techniques |
|---|---|---|---|---|---|
| Valieris et al. (2020) | Weakly-supervised | TCGA, Independent dataset | Breast | TCGA: AUC 0.80, Independent: AUC 0.70 | CNN for feature extraction, RNN for aggregation |
| Nero et al. (2022) | Weakly-supervised | Private dataset | Ovarian | Test: AUC 0.55 | CLAM |
| Bergstrom et al. (2023) | Weakly-supervised | TCGA, Independent validation cohorts | Breast, Ovarian | TCGA: AUC 0.81 (Breast), 0.76 (Ovarian) | Patch extraction, Transfer learning |
| Lazard et al. (2022) | Self-supervised and weakly-supervised | TCGA, Curie dataset | Breast | TCGA: AUC 0.71, Curie: AUC 0.86 | MoCo for feature extraction, Attention-based MIL, Bias control, Domain-specific augmentations |
| Bourgade et al. (2023) | Self-supervised and weakly-supervised | OvarIA cohort, TCGA | Ovarian (HGOC) | OvarIA: AUC 0.739 (5-fold), 0.681 (testing), TCGA: AUC 0.631 | CNN for tumor segmentation, BRCA classifier, MoCo for feature representation |
| Ahn et al. (2024) | Self-supervised and weakly-supervised | SEV cohort, TCGA-OV, SMC cohort | Ovarian | SEV: AUC 0.627, TCGA: AUC 0.602, SMC: AUC 0.593 | Contrastive self-supervised learning, CNN, MIL |
| Filiot et al. (2023) | Self-supervised and weakly-supervised | TCGA-OV, TCGA-BRCA | Ovarian, Breast | TCGA-OV: AUC 0.74, TCGA-BRCA: AUC 0.78 | ViT-B transformer, iBOT and MIM self-supervised methods, Transfer learning with MIL |

a significant challenge due to the high-dimensional nature of visual data. He et al. (2020) addressed this challenge by introducing Momentum Contrast (MoCo), a technique that constructs dynamic, large, and consistent dictionaries using contrastive loss. Their work demonstrated that MoCo effectively narrows the gap between unsupervised and supervised representations in computer vision tasks such as object detection and segmentation, using widely recognized datasets like PASCAL, VOC, and COCO. Building on these advancements, Chen et al. (2020) proposed a straightforward algorithm for contrastive learning for visual representation known as SimCLR. Their research highlighted the significance of data augmentation composition, the incorporation of a learnable nonlinear transformation between the representation and the contrastive loss, and the use of larger batch sizes (4k - 8k) combined with more training steps. These findings significantly enhanced model effectiveness, setting a new state-of-the-art on the ImageNet dataset. Later, researchers at Facebook AI Research introduced MoCov2 (Chen et al., 2020), which incorporated more aggressive data augmentation and a multi-layer perceptron projection head. This version demonstrated improved performance compared to the original MoCo and SimCLR, particularly on the ImageNet dataset. Importantly, MoCo v2 can process a large set of negative samples without requiring large training batches or powerful GPUs, making it feasible to run on a typical 8-GPU machine.

#### 2.3.2. *Applications of self-supervised method in computational pathology*

The impact of these self-supervised methods extends into computational pathology. Dehaene et al. (2020) leveraged MoCo v2 within a self-supervised learning framework to effectively close the gap between weakly-supervised and fully-supervised learning using histopathology images from the Camelyon16 dataset. This demonstrated the potential of self-supervised methods in enhancing diagnostic accuracy without the need for extensive annotated datasets. Li et al. (2021) introduced the Dual-stream Multiple Instance Learning Network (DS-MIL), which employs self-supervised contrastive learning which is derived from approaches similar to SimCLR for feature extraction. This model uses a novel dual-stream architecture that combines max-pooling and attention-based

aggregation to improve classification accuracy and localization performance. DS-MIL demonstrated superior performance on WSI classification tasks compared to previous MIL models, achieving high classification accuracy on datasets such as Camelyon16 and TCGA lung cancer. Chen et al. (2022) introduced a novel Vision Transformer (ViT) architecture called the Hierarchical Image Pyramid Transformer (HIPT) using the self-distillation with no labels (DINO) self-supervised method. HIPT leverages the hierarchical structure of WSIs with two levels of self-supervised learning to learn high-resolution image representations. Pre-trained on 10,678 gigapixel WSIs across 33 cancer types, HIPT outperforms state-of-the-art methods for cancer subtyping and survival prediction. Benchmarking on 9 slide-level tasks showed superior performance. Subsequently, they trained a MIL model using weak labels for a binary classification task on 1,008 WSIs of lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), achieving an impressive AUC of 0.952 ± 0.021.

#### *2.3.3. Self-supervised method with transfer learning for HRD detection*

However, in recent years, the application of self-supervised learning combined with weakly supervised techniques through transfer learning for HRD detection from breast and ovarian cancer WSIs has shown promising results as well. Lazard et al. (2022) used the self-supervised method MoCo to extract features from slides and incorporated the MIL model with attention-based aggregation for classification. Their approach achieved a mean AUC of 0.71 on the TCGA dataset and a mean AUC of 0.86 on the in-house Curie dataset for validation, for identifying HRD in breast cancer. They also addressed potential biases by incorporating strategic sampling and domain-specific augmentations to improve the robustness and generalization of the model. Furthermore, they identified several HRD-related morphological patterns, such as laminated fibrosis and clear tumor cells, highlighting the phenotypic impact of HRD. Bourgade et al. (2023) developed a deep learning framework for detecting BRCA mutations, which represent a significant proportion of HRD, in high-grade ovarian cancer (HGOC) using a CNN for tumor segmentation and feature extraction. They then used a BRCA classifier trained with attention-based MIL. The model also incorporated the self-supervised learning method MoCo for feature representation. Using the OvarIA cohort consisting of 867 HGOC patients, the BRCA classifier achieved an AUC of 0.739 in 5-fold cross-validation and an AUC of 0.681 on the internal testing set. When validated on 103 FFPE and H&E-stained slides from the TCGA dataset, the model achieved an AUC of 0.631. Ahn et al. (2024) developed PathoRiCH, which employs MIL models combined with CNN to predict HRD status from digital H&E slides. The study used contrastive self-supervised learning to enhance the model's performance. The datasets used included an in-house cohort from Yonsei Severance Hospital (SEV cohort) with 394 patients and 754 WSIs, the TCGA-OV dataset with 284 patients and 516 WSIs, and an external validation cohort from Samsung Medical Center (SMC cohort) with 136 patients and 136 WSIs. The PathoRiCH model achieved an AUC-ROC value of approximately 0.627 for the SEV cohort, 0.602 for the TCGA cohort, and 0.593 for the SMC cohort. Filiot et al. (2023) developed a ViT-B transformer model trained using iBOT and MIM self-supervised methods on 40 million pan-cancer WSIs. The resulting model, iBOT[ViT]PanCancer, was used to extract features, which were then employed in downstream HRD classification tasks using several MIL models via transfer learning. This approach achieved a mean AUC of 0.74 for TCGA-OV and 0.78 for TCGA-BRCA, demonstrating its effectiveness in HRD detection across different cancer types. Table 1 illustrates a summary of studies on HRD detection using histopathology images in breast and ovarian cancer.

## 3. Material and methods

In this section, we have discussed the materials and methods used for HR status prediction from breast and ovarian cancer WSIs.

### *3.1. Dataset*

In this study, we utilized both public and private datasets to ensure a comprehensive analysis of HR status prediction in breast and ovarian cancers. The public datasets were sourced from The Cancer Genome Atlas (TCGA), which provides extensive and high-quality digital pathology images widely used for research in computational pathology. We used the TCGA Ovarian Cancer (TCGA-OV) dataset, consisting of 1,394 WSIs from 558 patients, of which 312 patients had HR-deficient (HRD) status and 246 patients had HR-proficient (HRP) status. Additionally, the TCGA Breast Cancer (TCGA-BRCA) dataset was utilized, which includes 2,912 WSIs from 1,025 patients, with 204 patients having HRD status and 821 patients having HRP status. These datasets are instrumental for understanding the clinical and histopathological characteristics necessary for HR status identification in both breast and ovarian cancer.

To observe the impact of using different datasets, particularly for ovarian cancer, we incorporated a private dataset: the DIJON Ovarian Cancer (DIJON-OV) dataset. The DIJON-OV dataset was collected from the Georges-François Leclerc Regional Center for the Fight Against Cancer (CGFL) in Dijon, France. This dataset comprises 175 WSIs from 104 patients, among whom 42 patients had HRD status and 62 patients had HRP status. By integrating these public and private datasets,

our study benefits from a rich and varied data pool, which enables a thorough investigation of HR status especially in ovarian cancer. While we also conduct experiments on breast cancer data, our primary focus remains on ovarian cancer, as several studies have already been conducted for breast cancer.

### *3.2. Pre-processing*

WSIs are typically large, high-resolution, and multi-level digital images, often the size reaches in gigapixels. Therefore, handling or utilizing these high-resolution WSIs for computational analysis and model training is significantly challenging. Consequently, pre-processing these WSIs is essential before they can be used for deep learning model training. In this study, we utilized pre-processed WSIs generated by an existing pipeline developed by previous intern students for our experiments. The steps involved in this pre-existing workflow are as follows: Firstly, the maximum level, which has the highest resolution, of each WSI is extracted to retain the most useful and detailed cellular and tissue information. Following this, the maximum level of each WSI is divided into smaller tiles with a size of $256 \times 256$ pixels. This tiling approach makes the data more manageable and suitable for computational analysis. Next, intensity thresholding is applied to exclude tiles that predominantly contain the white background, as these tiles do not provide any useful cellular or tissue information. Specifically, the proportion of white pixels is computed in each tile. If a tile has more than 80% white pixels, it is removed from the dataset. This step effectively filters out background and unnecessary portions of the WSIs, ensuring that the dataset is refined and focused on areas with relevant tissue content. These pre-processing techniques significantly enhance the quality of the dataset, making it more suitable for sophisticated model training and analysis.

### *3.3. Fully supervised method*

In this study, we explored fully supervised technique to identify HR status in ovarian cancer WSIs. We employed several neural network architectures, including a CNN model composed of four convolutional layers, each followed by pooling layers to reduce dimensionality and retain essential features, as well as ResNet34 and ResNet50 (He et al., 2016). Training these neural network architectures using WSIs is challenging as it requires significant memory and computational resources. Therefore, we used smaller tiles extracted from the WSIs, as detailed in Section 3.2 on pre-processing.

However, generating tiles from WSIs introduces a new challenge. The fully supervised method requires labels for each tile, but we only have slide-levels or global labels. Consequently, manually annotating each tile is quite impossible since pathologists cannot determine which tiles are HR-deficient and which are HR-proficient without additional genetic tests. To address this issue, we generated pseudo-labels for each tile based on a hypothesis. This hypothesis asserts that all tiles derived from a globally HR-deficient slide should be labeled as HRD, even though some tiles may not truly be HR-deficient. Conversely, for slides that are globally labeled as HR-proficient, all generated tiles will be accurately labeled as HRP.

Mathematically, the hypothesis can be expressed as follows:

$$\text{Label}(T_i) = \begin{cases} \text{HRD} & \text{if Label}(S) = \text{HRD} \\ \text{HRP} & \text{if Label}(S) = \text{HRP} \end{cases}$$

Where $T_i$ represents each tile and $S$ represents the entire slide.

Subsequent to creating pseudo labels, we used annotated tiles obtained from HRD and HRP globally labeled slides to train the neural network models. This approach allowed us to leverage the large dataset effectively while adhering to the fully supervised learning framework.

#### *3.3.1. Training and experiment*

In this study, we conducted multiple experiments using various datasets to train different neural network architectures, specifically CNN, ResNet34, and ResNet50. Given our primary focus on identifying HR status in ovarian cancer, we specifically began our experiments with the DIJON-OV dataset.

For the DIJON-OV experiment, we split the dataset into 80% for training and 20% for validation. To confirm the integrity of the dataset, we ensured that all tiles generated from a WSI remained together and were not replicated across different WSIs. This approach prevents data leakage and ensures robust model evaluation. Next, we utilized the training dataset for hyperparameter tuning to determine the optimal learning rate for each of the three neural network models. Finding the best learning rate is crucial because it directly affects the convergence speed and stability of the training process, ultimately minimizing the loss during training. By identifying the optimal learning rate, we ensured that the models could learn efficiently and effectively, reducing the likelihood of overshooting minima or getting stuck in suboptimal points. In addition, we employed the Adam optimizer in conjunction with the binary cross-entropy loss function, and we set the batch size to 34. We then proceeded to train the CNN, ResNet34, and ResNet50 models using the training dataset and the defined hyperparameters. To validate the performance of these models, we used the split validation dataset. This validation step is essential to assess the generalizability and effectiveness of the trained models in identifying HR status in ovarian cancer. This experiment was conducted using the computing server of the Laboratory for Research on Learning and Development at the University of Burgundy. The server was configured with an

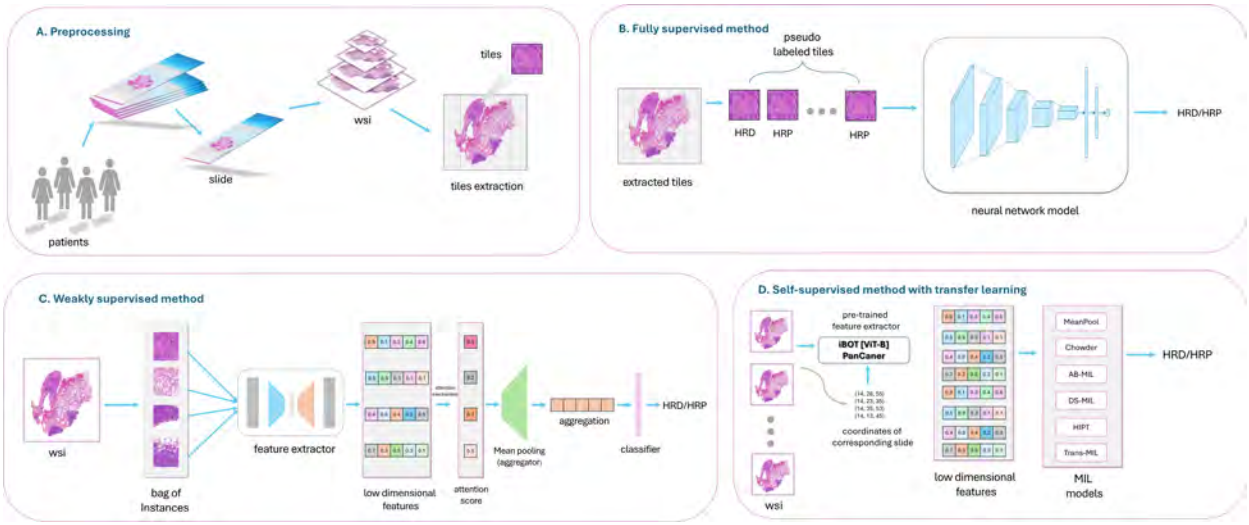



Figure 1: Block diagram of (a) preprocessing, (b) fully supervised method, (c) weakly supervised method, and (d) self-supervised method combined with transfer learning.

NVIDIA RTX A6000 GPU with 48 GB of memory, utilizing CUDA version 12.0 and cuDNN version 8.0 for optimized deep learning framework performance. The system featured an AMD EPYC 7343 processor operating at 3.2 GHz (2P, 16C/P) and was equipped with 256 GB of system RAM.

Next, we extended our experiments to the TCGA-OV dataset, following the same 80:20 ratio for the data split for training and validation as in our previous experiment. Given the substantial size of the TCGA-OV dataset, which includes a large number of slides and corresponding tiles, we employed Horovod, a distributed training framework developed by Uber, to expedite the training process (Sergeev and Del Balso, 2018). Horovod enables data parallelism, where each GPU gets a subset of the data and computes gradients on its subset (Sergeev and Del Balso, 2018). These gradients are then averaged and used to update the model parameters (Sergeev and Del Balso, 2018). This technique significantly reduced the training time, almost by a factor of four compared to training without Horovod. For this phase, we increased the batch size to 64, while keeping other hyperparameters consistent with those used in the DIJON-OV experiments. We performed these experiments on the CCUB server, the computing center of the University of Burgundy, configured with the following specifications: 32 cores, 256 GB memory, an AMD EPYC 7343 @ 3.2 GHz (2P, 16C/P) combined with 2 Tesla A100 GPUs with 40 GB of RAM. This high-performance setup provided the necessary computational resources to handle the extensive data and complex calculations involved in our deep-learning tasks.

As our primary focus was to analyze the performance of HR status prediction in ovarian cancer, we did not conduct a fully supervised experiment with the TCGA-BRCA dataset. Instead, we concentrated on leveraging more advanced techniques for analyzing HRD prediction results in TCGA-BRCA during the later part of our study, which allowed us to utilize our resources more effectively. However, this comprehensive training strategy enabled us to effectively utilize large-scale datasets and optimize the performance of our CNN models for HR status identification in ovarian cancer WSIs. The integration of advanced training techniques and hyperparameter tuning ensured that our models were both accurate and efficient, paving the way for reliable HR status classification.

### *3.4. Weakly supervised method*

In recent years, weakly supervised methods, such as MIL, have gained significant traction in histopathology classification tasks due to their ability to achieve state-of-the-art results. As a result, we also explored this technique for predicting HR status from ovarian cancer WSIs. For our experiments, we utilized the DIJON-OV dataset, which contains a comprehensive collection of ovarian cancer samples.

#### *3.4.1. Training and experiment*

In our experiment utilizing the weakly supervised method with the DIJON-OV dataset, the process began with a preprocessing phase where tiles, also known as instances, were generated from all the WSIs of DIJON-OV. These instances were created using the standardized pre-existing preprocessing pipeline discussed in section 3.2 to ensure consistency and quality across all samples. Each slide was represented as a bag containing its corresponding instances, adhering to the feature extractor. Next, we employed a denoising autoencoder to extract features from the instances within each bag. This autoencoder consisted of four residual blocks in the encoder and four transposed convolutional layers in the

decoder. Each residual block in the encoder included two convolutional layers with kernel sizes of 3x3, batch normalization, and ReLU activation, with the second convolutional layer having a stride of 2 for downsampling. The decoder mirrored this structure with transposed convolutional layers to upsample the encoded representation. Additionally, a Gaussian noise layer was added before the encoding process to enable the model to learn robust features by denoising the input images. A denoising autoencoder is a type of neural network designed to learn important features of input data while also being robust to noise, which helps in reducing the dimensionality of the feature space while retaining essential information. The extracted features were then projected into a lower-dimensional space, making them more manageable for subsequent processing. The extracted lower-dimensional features were then used for training the MIL model. Within the MIL model, these features were processed through an attention mechanism to identify the most important features of the corresponding tiles. This mechanism assigns an attention weight to each feature based on its significance, effectively highlighting the most informative parts of the data. Attention mechanisms are particularly useful in scenarios where certain features play a crucial role in the classification task, allowing the model to focus on these key elements. Subsequently, we employed a MIL pooling aggregator to combine the high-weighted features from the instances within each bag. The MIL pooling aggregator pools the attention-weighted features, leading to a comprehensive representation of the slide-level data. This aggregated representation was then used to make the final global-level classification of HR status. In this experiment, we trained the MIL model using the standard 5-fold cross-validation (CV) method. The hyperparameters were set as follows: a learning rate of 0.005, the Adam optimizer, binary cross-entropy as the loss function, and a batch size of 1 to facilitate bag-level training. This approach ensures robust evaluation and effective learning from the available data, leveraging the attention mechanism and pooling strategies to enhance the model's performance.

In this experiment, we only used the attention-based MIL model. However, there are various types of MIL models such as CLAM, mean pool, and TransMIL. We utilized these models with a pre-trained self-supervised model used for extracting features from both ovarian and cancer WSIs to identify the HRD status.

### *3.5. Self-supervised method with transfer learning*

In our study, we utilized a ViT-B architecture trained with self-supervised learning techniques such as iBOT (Zhou et al., 2022) combined with MIM (Xie et al., 2022) on a dataset of 40 million pan-cancer WSIs developed by Filiot et al. (2023). iBOT focuses on instance-aware learning, enabling the model to bootstrap its learning process by leveraging the relationships between instances within the data. MIM, on the other hand, involves masking parts of the input image and training the model to predict the missing pieces, thereby encouraging the model to learn contextual representations. This model, referred to as iBOT[ViT-B]PanCancer, was used to extract features from WSIs for our experiments. The use of the pre-trained iBOT[ViT-B]PanCancer model for feature extraction highlights the application of transfer learning in our study. By leveraging a model that was pre-trained on a large dataset of pan-cancer WSIs, we could effectively transfer the learned representations to our specific task, thereby enhancing the efficiency and accuracy of feature extraction from our histopathological slides. We followed the instructions and maintained the folder structure provided by the GitHub repository from (https://github.com/owkin/HistoSSLscaling) by Filiot et al. (2023). The feature extraction process using this pre-trained model involves several important steps. First, we extracted all coordinates from the highest resolution level of the WSIs. These coordinates were then filtered to remove those that did not correspond to useful tissue information, ensuring that only regions with relevant tissue were retained. This step is crucial for identifying regions containing relevant tissue information while excluding areas without useful content such as a white background. This process helps significantly reduce the computational burden and the time required for feature extraction. After successfully extracting useful coordinates which only belong to tissue regions from slides, we used these coordinates in conjunction with the pre-trained iBOT[ViT-B]PanCancer model to extract features from the slides.

The extracted features from each slide were then utilized for downstream classification tasks using various MIL models. These MIL models included Mean-pool MIL, Chowder MIL, AB-MIL, DS-MIL, and Trans-MIL, each contributing to the slide-based classification. Mean-pool MIL aggregates features by averaging them, providing a simple yet effective summary. Chowder MIL employs a more sophisticated aggregation method to capture complex patterns. AB-MIL and DS-MIL offer attention-based mechanisms to weigh the importance of different instances, enhancing the model's focus on critical regions. Finally, Trans-MIL leverages transformer-based architectures for capturing long-range dependencies within the slide. Together, these techniques enabled robust and accurate classification of histopathological slides, demonstrating the efficacy of self-supervised learning methods in this domain.

#### *3.5.1. Training and experiment*

We believe that this approach is one of the robust and generalized approaches to perform the experiments for HR status prediction from both breast and ovarian cancer. Therefore, we started our experiments with both the public datasets of breast and ovarian cancer, such as

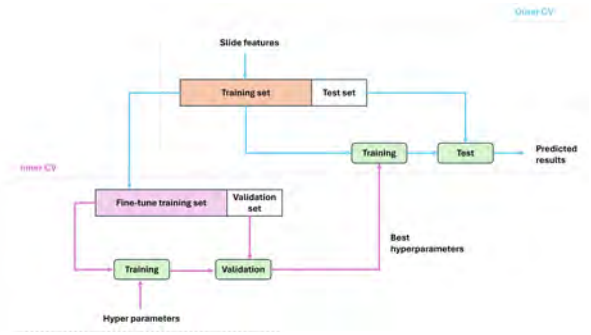



Figure 2: Block diagram of nested cross-validation.

TCGA-OV and TCGA-BRCA, as this dataset contains a huge number of WSIs which will give more generalized performances. We used the same techniques for the classification task for both datasets. We used a nested CV technique shown in Figure 2 for training and testing the MIL models for the classification tasks. A nested CV consists of two parts: the outer CV and the inner CV. The inner CV is used for tuning hyperparameters, while the outer CV is used for making training and testing predictions using the optimal hyperparameters determined by the inner CV. To perform the nested CV, we defined the number of repetitions and splits for both the inner and outer CV. We chose to perform 1 iteration and 5 splits for both the inner and outer CV processes.

In each split of the outer CV, all the features are divided into five folds, one of them is used as the test set while others are used as the training set. The training set from each split of the outer CV is then further divided into five folds by the inner CV, where one fold is used as the validation set and the other folds are used as the fine-tuning training set. This fine-tuning training set is employed to train the model using various combinations of hyperparameters. In this experiment, we focused on tuning the learning rate and decay rate, selecting two learning rates (0.001 and 0.0001) and two decay rates (0 and 0.0001) proposed by Filiot et al. (2023) for fine-tuning. Once the model is trained with each combination of learning rate and decay rate, it is validated using the validation set. This process is repeated for each split of the inner CV. The combination of learning rate and decay rate that yields the best validation result is selected as the optimal hyperparameter set. These optimal hyperparameters are then used to train the model on the training set of the outer CV. After the training phase, the model is tested using the test set. This process is repeated for each split in the outer CV, ensuring a thorough evaluation of the model's performance. It is important to mention that we used stratified and patient split modes to ensure a balanced distribution of HRD and HRP-labeled slides in each fold during cross-validation. This approach helps maintain an equal representation of both classes, providing more reliable and accurate model evaluations. Additionally, we used the Adam optimizer and the binary cross-entropy loss function, setting the batch size to 16 for our training process.

Although the experiment was limited to publicly available datasets for breast and ovarian cancer, we also plan to conduct experiments with private datasets. This will help us to understand the impact of the dataset on the identification performance of HR status, specifically in ovarian cancer.

## 4. Results

After conducting several experiments with various training strategies and models on several datasets, we obtained comprehensive results and performance metrics. These metrics illustrate the capability of the models in predicting HRD in both breast cancer and ovarian cancer. In this study, we utilize several performance metrics to evaluate the effectiveness of the model. The Area Under the Curve (AUC) measures the ability of a model to distinguish between positive and negative classes, with a higher AUC indicating better discrimination. Accuracy assesses the overall correctness of the model by calculating the proportion of correct predictions (both true positives and true negatives) out of the total number of cases examined, providing a straightforward measure of general performance. Sensitivity (also known as the True Positive Rate or Recall) measures the proportion of actual positive cases that the model correctly identifies, highlighting the model's capability to detect positive instances accurately. Specificity (also known as the True Negative Rate) quantifies the proportion of actual negative cases that the model correctly identifies, indicating the model's ability to avoid misclassifying negative instances as positive. Finally, the F1-score provides a balance between precision and recall, highlighting the model's accuracy in both identifying positive instances and not mislabeling negative instances as positive. It is the harmonic mean of precision and recall, giving a single metric that considers both false positives and false negatives.

### *4.1. Experimental results on DIJON-OV*

The performance metrics of various models on the DIJON-OV dataset are shown in Table 2. For the fully supervised technique, all models struggled to classify the positive samples (HRD). The CNN model had the lowest performance with an AUC of 0.47 and an accuracy of 0.51, along with suboptimal sensitivity and specificity of 0.32 and 0.63, respectively, indicating a potential bias toward negative samples. ResNet34 and ResNet50 showed slightly improved performance with the highest specificity of 0.65 reported for ResNet34 and the best AUC of 0.55 for ResNet50. Overall, ResNet34 provided a balanced performance among three neural network models. It is important to note that these results were obtained using single-split validation

Table 2: Performance metrics of various models on the DIJON-OV dataset (mean ± standard deviation for some values)

| Dataset | Technique | Feature Extractor | Model | AUC | Accuracy | Sensitivity | Specificity | F1-Score |
|---|---|---|---|---|---|---|---|---|
| DIJON-OV | Fully Supervised | | CNN | 0.47 | 0.51 | 0.32 | 0.63 | 0.36 |
| | | | ResNet34 | 0.54 | 0.55 | 0.40 | **0.65** | 0.40 |
| | | | ResNet50 | **0.55** | 0.54 | 0.37 | 0.64 | 0.37 |
| | Weakly Supervised | Auto Encoder | AB-MIL | 0.53 ± 0.04 | **0.58 ± 0.05** | **0.43 ± 0.04** | 0.61 ± 0.15 | **0.46 ± 0.27** |

Table 3: Performance metrics of various models on TCGA-OV dataset (mean ± standard deviation for some values)

| Dataset | Technique | Feature Extractor | Model | AUC | Accuracy | Sensitivity | Specificity | F1-Score |
|---|---|---|---|---|---|---|---|---|
| TCGA-OV | Fully Supervised | | CNN | 0.50 | **0.62** | 1.00 | 0.00 | **0.77** |
| | | | ResNet34 | 0.58 | 0.57 | 0.60 | 0.51 | 0.63 |
| | | | ResNet50 | 0.59 | 0.58 | 0.67 | 0.44 | 0.66 |
| | Self Supervised | iBOT[ViT-B] PanCancer | Mean-Pool | 0.65 ± 0.50 | 0.61 ± 0.04 | 0.65 ± 0.10 | 0.54 ± 0.11 | 0.65 ± 0.06 |
| | | | Chowder | **0.66 ± 0.06** | 0.60 ± 0.04 | 0.59 ± 0.09 | **0.63 ± 0.14** | 0.62 ± 0.05 |
| | | | AB-MIL | 0.65 ± 0.04 | 0.61 ± 0.04 | 0.76 ± 0.10 | 0.41 ± 0.15 | 0.69 ± 0.05 |
| | | | DS-MIL | 0.63 ± 0.03 | 0.61 ± 0.02 | **0.77 ± 0.13** | 0.39 ± 0.17 | 0.69 ± 0.05 |
| | | | HIPT | 0.64 ± 0.05 | 0.61 ± 0.05 | **0.77 ± 0.12** | 0.40 ± 0.19 | 0.69 ± 0.05 |
| | | | Trans-MIL | **0.66 ± 0.03** | 0.61 ± 0.05 | 0.73 ± 0.25 | 0.47 ± 0.26 | 0.66 ± 0.12 |

Table 4: Performance metrics of various models on TCGA-BRCA dataset (mean ± standard deviation)

| Dataset | Technique | Feature Extractor | Model | AUC | Accuracy | Sensitivity | Specificity | F1-Score |
|---|---|---|---|---|---|---|---|---|
| TCGA-BRCA | Self Supervised | iBOT[ViT-B] PanCancer | Mean-Pool | 0.76 ± 0.04 | 0.81 ± 0.01 | 0.25 ± 0.09 | **0.96 ± 0.02** | 0.34 ± 0.08 |
| | | | Chowder | 0.75 ± 0.06 | 0.80 ± 0.02 | 0.27 ± 0.16 | 0.93 ± 0.07 | 0.31 ± 0.18 |
| | | | AB-MIL | **0.77 ± 0.06** | 0.81 ± 0.02 | 0.25 ± 0.08 | 0.95 ± 0.01 | 0.33 ± 0.08 |
| | | | DS-MIL | 0.74 ± 0.04 | **0.82 ± 0.02** | **0.30 ± 0.07** | 0.94 ± 0.02 | **0.39 ± 0.07** |
| | | | HIPT | 0.75 ± 0.05 | 0.81 ± 0.02 | 0.25 ± 0.14 | 0.95 ± 0.09 | 0.32 ± 0.16 |
| | | | Trans-MIL | 0.76 ± 0.03 | 0.81 ± 0.02 | 0.21 ± 0.10 | **0.96 ± 0.02** | 0.30 ± 0.10 |

data. We computed all metrics for this experiment by considering the approximate mean values of the last ten epochs of validation results.

In contrast, the weakly supervised method, AB-MIL with Autoencoder, demonstrated a better sensitivity. This method showed improved sensitivity, with a mean sensitivity of 0.43 and a standard deviation of 0.04, which is higher than neural network models. Although the specificity (0.61 ± 0.15) was almost close to the ResNet models, the results indicate that AB-MIL can provide a more generalized performance for classifying both HRD and HRP classes, even with slide-based classification. Furthermore, AB-MIL achieved the highest F1-score at 0.46 ± 0.27. These results were obtained using a standard 5-fold CV.

However, we hypothesize that using more advanced techniques for feature extraction or employing pre-trained models with a large amount of digital pathology data could enhance the performance of AB-MIL. This hypothesis was tested in the TCGA-OV experiment.

### *4.2. Experimental results on TCGA-OV*

The performance metrics of various models on the TCGA-OV dataset are shown in Table 3. In the fully supervised method, the CNN model demonstrated a significant bias towards sensitivity, predicting all samples as HRD cases. Consequently, the model completely failed to identify any HRP cases, resulting in a specificity of 0.00. This extreme bias towards sensitivity (1.00) led to a deceptively high F1-score (0.77). Due to this bias, the CNN model lacked the ability to generalize predictions and effectively distinguish between the two classes. The ResNet34 model showed improvements over CNN in terms of model generalization, suggesting it could better discriminate between positive and negative cases. However, this model did not exhibit very high sensitivity and specificity, its overall performance was more balanced with an AUC of 0.58, sensitivity of 0.60, and specificity of 0.51. It still fell short of being ideal. ResNet50 offered further improvements, increasing the sensitivity to 0.67 compared to ResNet34. However, it did so at the cost of specificity (0.44). Although the improvements in ResNet50 were not highly significant, it achieved a slightly higher AUC of 0.59 and F1-score of 0.66 among all the neural network models evaluated, indicating a marginally better overall performance. It is important to note that these results were obtained using single-split validation data. We computed

all metrics for this experiment by considering the approximate mean values of the last five epochs of validation results.

The self-supervised with transfer learning methods, utilizing the iBOT[ViT-B] PanCancer as a feature extractor, demonstrated varied performances across different models. The Mean-Pooling approach showed a balanced performance, achieving a moderate mean AUC of 0.65 with a 0.50 standard deviation and a mean F1-score of 0.65 with a standard deviation of 0.06. Its sensitivity (0.65 ± 0.10) and specificity (0.54 ± 0.11) were relatively well-balanced, indicating a good ability to generalize predictions and distinguish between HRD and HRP cases effectively. The Chowder model also performed well, achieving a high mean AUC of 0.66 with a 0.06 standard deviation. It provided the least sensitivity (0.59 ± 0.09) and the highest specificity score (0.63 ± 0.14), suggesting it is more robust towards HRP cases. The AB-MIL model demonstrated a strong sensitivity of 0.76 ± 0.10, effectively identifying HRD cases. However, this came at the cost of a lower specificity of 0.41 ± 0.15, indicating some difficulty in correctly identifying HRP cases. Despite this, the model achieved a high F1-score (0.69 ± 0.05), reflecting its robustness in detecting positive cases. The DS-MIL model exhibited a similar pattern to AB-MIL, with high sensitivity (0.77 ± 0.13) but lower specificity (0.39 ± 0.17). This imbalance led to an F1-score (0.69 ± 0.05), highlighting the model's effectiveness in identifying HRD cases, though it struggled with HRP cases. The HIPT model also showed strong sensitivity (0.77 ± 0.12), similar to DS-MIL, with a balanced overall performance. The high sensitivity and moderate specificity (0.40 ± 0.19) resulted in a high F1-score (0.69 ± 0.05), indicating the model's robustness in HRD detection. The Trans-MIL model achieved the highest AUC of 0.66 ± 0.03 (although the mean AUC is similar to Chowder, the standard deviation is less than Chowder) among the self-supervised techniques, indicating strong discriminative power. It balanced sensitivity (0.73 ± 0.25) and specificity (0.47 ± 0.26) well, resulting in a high F1 score (0.66 ± 0.12) compared to Chowder MIL. This balanced performance suggests that Trans-MIL is particularly effective in predicting HRD status, making it one of the standout MIL models. These results were obtained using the nested CV.

Overall, the self-supervised with transfer learning techniques demonstrated strong performance, with models like Chowder and Trans-MIL showing particularly balanced and effective results. These methods exhibited a good ability to generalize predictions and distinguish between HRD and HRP cases, underscoring their potential in HR status prediction from digital histopathology images in ovarian cancer.

### *4.3. Experimental results on TCGA-BRCA*

As we already mentioned, the objective of experimenting with HRD identification from breast cancer is to learn about the performance capability of the pipelines used. As HRD detection from breast cancer has already provided outstanding results, we hypothesise that the pipeline we have used will work well for HRD detection from breast cancer. To utilize our resources properly we just used the self-supervised method combined with transfer learning here as this approach provided more robust and generalized results in our previous experiments with ovarian cancer data. The experimental results for TCGA-BRCA are shown in Table 4.

From the table, it is evident that the AB-MIL model achieves the highest mean of AUC 0.77 with 0.06 standard deviation, although this model, like others, exhibits a notable disparity between sensitivity and specificity, with very high specificity and low sensitivity. This trend is consistent across all models evaluated. One plausible explanation for this pattern is the significant class imbalance in the TCGA-BRCA dataset, where the number of HRP samples is substantially higher, almost four times, than the HRD samples. This imbalance likely contributes to the observed lower sensitivity and subsequently impacts the F1-score, as the sensitivity is a critical component of the F1-score. Despite AB-MIL's superior AUC, the DS-MIL model demonstrates the highest accuracy (0.82 ± 0.02), sensitivity (0.30 ± 0.07), and F1-score (0.39 ± 0.07), indicating a more balanced performance across different metrics. Although DS-MIL's specificity is not the highest, it offers a more generalized ability to distinguish between HRD and HRP classes, making it a potentially more reliable model for practical applications. Other models such as Mean-Pool, Chowder, HIPT, and Trans-MIL also show competitive performance, with results closely aligning with those of AB-MIL and DS-MIL. This suggests that while AB-MIL and DS-MIL have certain advantages, the other models are also viable options depending on the specific requirements and constraints of the analysis. These results were also obtained using nested CV.

In summary, the experiment underscores the importance of considering multiple performance metrics when evaluating model efficacy, particularly in the context of imbalanced datasets. The results highlight the nuanced trade-offs between sensitivity, specificity, and overall accuracy, providing valuable insights for optimizing HRD detection pipelines in breast and ovarian cancer research.

## 5. Discussion

The experiments conducted using various learning techniques and models with both public and private data on breast and ovarian cancer have revealed several valuable insights into the prediction of HRD biomarkers.

One of the key findings is that predicting HRD biomarkers is more reliable and easier in breast cancer compared to ovarian cancer. This difference is likely due to the differing complexity in the morphological patterns of the tissues, with ovarian cancer potentially being more complex. This complexity can make it more challenging for models to accurately predict HRD in ovarian cancer. Additionally, our comparison of the TCGA-BRCA and TCGA-OV datasets is robust because both use similar protocols for data collection, have similar patient demographics, and use consistent labeling practices, unlike the DIJON-OV dataset.

Our findings also suggest that the performance of HRD biomarker prediction is highly contingent on class balance of the dataset. A significant observation is the tendency of models commonly used for histopathology classification to exhibit bias towards the more prevalent class. For instance, in the TCGA-BRCA experiment, all models demonstrated high specificity scores due to the HRP samples outnumbering HRD samples by a factor of four, highlighting the issue of dataset imbalance. This scenario reveals a need for exploring various data augmentation and scaling techniques to effectively address the prevalent issue of dataset imbalance.

In terms of the methods and models employed for HRD status prediction, our experiments suggest that combining self-supervised and weakly supervised methods through transfer learning can yield more robust performance compared to relying solely on fully supervised or weakly supervised methods. The fully supervised approach, while effective, demands extensive annotation at the tile level, which is complex and time-consuming. On the other hand, weakly supervised methods, such as MIL, are suitable for classification tasks as predictions can be made at the slide level. However, these methods face limitations in feature extraction, often relying on neural network models like ResNet50 pre-trained on ImageNet weights, which are not specifically optimized for digital pathology data. Each MIL model has distinct strengths, and performance can vary depending on task complexity and dataset characteristics. Advanced models such as HIPT and Trans-MIL potentially offer improved performance but demand greater computational resources and time, making them less practical for broader applications compared to more general models like AB-MIL, Chowder, or Mean-Pool. Given these considerations, a strategic approach involves utilizing self-supervised methods such as iBOT, MIM, MoCo, and DINO. These methods allow for training models on vast amounts of WSI data across different cancer subtypes without the need for labeled data. The resulting pre-trained models can then be employed for feature extraction, capturing more relevant and impactful features for slide-level classification.

Nevertheless, our conclusions are tentative, relying primarily on the analysis of TCGA datasets with slide-based levels. There is a compelling need for additional research using diverse datasets, such as DIJON-OV, to evaluate the performance of slide-based classification via self-supervised learning combined with transfer learning. Such investigations could provide deeper insights into HRD prediction for ovarian cancer WSIs.

An exciting avenue for future research could involve the application of foundation models, such as GigaPath recently published by Microsoft Health Futures (Xu et al., 2024). This novel approach utilizes a new and advanced vision transformer architecture for handling gigapixel pathology images, leveraging a diverse real-world cancer patient dataset. GigaPath aims to lay a foundation for AI in cancer pathology and could be instrumental in improving HRD prediction for both ovarian and breast cancer. By integrating these advanced models, researchers can potentially achieve more accurate and robust performance, addressing current limitations and pushing the boundaries of HRD biomarker prediction.

While significant progress has been made, our research underscores the need for balanced datasets, the strategic combination of learning methods, and the exploration of advanced foundation models to enhance HRD biomarker prediction. Future research should focus on these areas to develop more reliable, accurate, and generalizable models for HRD status prediction in various cancer types.

## 6. Conclusions

Our study highlights challenges and complexities in predicting HRD in ovarian and breast cancer. Through several experiments using diverse datasets and methodologies, we have found that multiple factors, including dataset characteristics, class balance, and the choice of training methods and models, heavily influence HRD prediction performance. One key insight is that predicting HRD in ovarian cancer is more challenging than in breast cancer, likely due to the more complex morphological patterns in ovarian cancer tissues. This finding emphasizes the need for targeted research on ovarian cancer to develop more effective predictive models. To tackle these challenges, we recommend using robust techniques that focus on exploratory data analysis, such as data augmentation and scaling. Furthermore, advanced modeling techniques, including foundation models, could potentially enhance the accuracy and robustness of HRD predictions. In summary, our research underscores the critical need for ongoing advancements in dataset management and modeling approaches to enhance the precision of HRD biomarker predictions. These improvements are essential for personalizing treatment plans for cancer patients. Our study provides a foundation for future research aimed at refining predictive capabilities for HRD in both breast

and ovarian cancers, with a particular focus on overcoming the unique challenges of ovarian cancer.

## Acknowledgments

I extend my heartfelt gratitude to Manon Ansart and Patrick Bard for their invaluable guidance and supervision throughout this research project, as well as for hosting me at the LEAD-CNRS, University of Burgundy, France. I am also deeply grateful to our project partner, the Centre Georges-François Leclerc in Dijon, France, for their generous provision of the dataset that was essential to this study.

## References

Abkevich, V., Timms, K.M., Hennessy, B.T., Potter, J., Carey, M.S., Meyer, L.A., Smith-McCune, K., Broaddus, R., Lu, K.H., Chen, J., Tran, T.V., Williams, D., Iliev, D., Jammulapati, S., FitzGerald, L.M., Krivak, T., DeLoia, J.A., Gutin, A., Mills, G.B., Lanchbury, J.S., 2012. Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. British Journal of Cancer 107, 1776–1782. URL: https://www.nature.com/articles/bjc2012451, doi:10.1038/bjc.2012.451.

Ahn, B., Moon, D., Kim, H.S., Lee, C., Cho, N.H., Choi, H.K., Kim, D., Lee, J.Y., Nam, E.J., Won, D., An, H.J., Kwon, S.Y., Shin, S.J., Jung, H.R., Kwon, D., Park, H., Kim, M., Cha, Y.J., Park, H., Lee, Y., Noh, S., Lee, Y.M., Choi, S.E., Kim, J.M., Sung, S.H., Park, E., 2024. Histopathologic image–based deep learning classifier for predicting platinum-based treatment responses in high-grade serous ovarian cancer. Nature Communications 15, 4253. URL: https://www.nature.com/articles/s41467-024-48667-6, doi:10.1038/s41467-024-48667-6. publisher: Nature Publishing Group.

Anaya, J., Sidhom, J.W., Mahmood, F., Baras, A.S., 2024. Multiple-instance learning of somatic mutations for the classification of tumour type and the prediction of microsatellite status. Nature Biomedical Engineering 8, 57–67. URL: https://www.nature.com/articles/s41551-023-01120-3, doi:10.1038/s41551-023-01120-3. publisher: Nature Publishing Group.

Bergstrom, E.N., Abbasi, A., Díaz-Gay, M., Galland, L., Lippman, S.M., Ladoire, S., Alexandrov, L.B., 2023. Deep learning predicts HRD and platinum response from histology slides in breast and ovarian cancer. URL: https://www.medrxiv.org/content/10.1101/2023.02.23.23285869v1, doi:10.1101/2023.02.23.23285869. iSSN: 2328-5869 Pages: 2023.02.23.23285869.

Birkbak, N.J., Wang, Z.C., Kim, J.Y., Eklund, A.C., Li, Q., Tian, R., Bowman-Colin, C., Li, Y., Greene-Colozzi, A., Iglehart, J.D., Tung, N., Ryan, P.D., Garber, J.E., Silver, D.P., Szallasi, Z., Richardson, A.L., 2012. Telomeric Allelic Imbalance Indicates Defective DNA Repair and Sensitivity to DNA-Damaging Agents. Cancer Discovery 2, 366–375. URL: https://doi.org/10.1158/2159-8290.CD-11-0206, doi:10.1158/2159-8290.CD-11-0206.

Bourgade, R., Rabilloud, N., Perennec, T., Pécot, T., Garrec, C., Guédon, A.F., Delnatte, C., Bézieau, S., Lespagnol, A., de Tayrac, M., Henno, S., Sagan, C., Toquet, C., Mosnier, J.F., Kammerer-Jacquet, S.F., Loussouarn, D., 2023. Deep Learning for Detecting BRCA Mutations in High-Grade Ovarian Cancer Based on an Innovative Tumor Segmentation Method From Whole Slide Images. Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc 36, 100304. doi:10.1016/j.modpat.2023.100304.

Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine 25, 1301–1309. URL: https://www.nature.com/articles/s41591-019-0508-1, doi:10.1038/s41591-019-0508-1. publisher: Nature Publishing Group.

Chan, C.W.H., Law, B.M.H., So, W.K.W., Chow, K.M., Waye, M.M.Y., 2017. Novel Strategies on Personalized Medicine for Breast Cancer Treatment: An Update. International Journal of Molecular Sciences 18, 2423. doi:10.3390/ijms18112423.

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: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16123–16134. URL: https://ieeexplore.ieee.org/document/9880275, doi:10.1109/CVPR52688.2022.01567. iSSN: 2575-7075.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A Simple Framework for Contrastive Learning of Visual Representations, in: Proceedings of the 37th International Conference on Machine Learning, PMLR. pp. 1597–1607. URL: https://proceedings.mlr.press/v119/chen20j.html. iSSN: 2640-3498.

Chopra, N., Tovey, H., Pearson, A., Cutts, R., Toms, C., Proszek, P., Hubank, M., Dowsett, M., Dodson, A., Daley, F., Kriplani, D., Gevensleben, H., Davies, H.R., Degasperi, A., Roylance, R., Chan, S., Tutt, A., Skene, A., Evans, A., Bliss, J.M., Nik-Zainal, S., Turner, N.C., 2020. Homologous recombination DNA repair deficiency and PARP inhibition activity in primary triple negative breast cancer. Nature Communications 11, 2662. doi:10.1038/s41467-020-16142-7.

Cireşan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J., 2013. Mitosis detection in breast cancer histology images with deep neural networks. Medical image computing and computer-assisted intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention 16, 411–418. doi:10.1007/978-3-642-40763-5_51.

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. URL: https://www.nature.com/articles/s41591-018-0177-5, doi:10.1038/s41591-018-0177-5. publisher: Nature Publishing Group.

Courtiol, P., Tramel, E.W., Sanselme, M., Wainrib, G., 2020. Classification and Disease Localization in Histopathology Using Only Global Labels: A Weakly-Supervised Approach. URL: http://arxiv.org/abs/1802.02212, doi:10.48550/arXiv.1802.02212. arXiv:1802.02212 [cs, stat].

Davies, H., Glodzik, D., Morganella, S., Yates, L.R., Staaf, J., Zou, X., Ramakrishna, M., Martin, S., Boyault, S., Sieuwerts, A.M., Simpson, P.T., King, T.A., Raine, K., Eyfjord, J.E., Kong, G., Borg, , Birney, E., Stunnenberg, H.G., van de Vijver, M.J., Børresen-Dale, A.L., Martens, J.W.M., Span, P.N., Lakhani, S.R., Vincent-Salomon, A., Sotiriou, C., Tutt, A., Thompson, A.M., Van Laere, S., Richardson, A.L., Viari, A., Campbell, P.J., Stratton, M.R., Nik-Zainal, S., 2017. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nature Medicine 23, 517–525. doi:10.1038/nm.4292.

Dehaene, O., Camara, A., Moindrot, O., de Lavergne, A., Courtiol, P., 2020. Self-Supervision Closes the Gap Between Weak and Strong Supervision in Histology. URL: http://arxiv.org/abs/2012.03583, doi:10.48550/arXiv.2012.03583. arXiv:2012.03583 [cs, eess].

Deng, S., Zhang, X., Yan, W., Chang, E.I.C., Fan, Y., Lai, M., Xu, Y., 2020. Deep learning in digital pathology image analysis: a survey. Frontiers of Medicine 14, 470–487. doi:10.1007/s11684-020-0782-9.

Diao, J.A., Wang, J.K., Chui, W.F., Mountain, V., Gullapally, S.C., Srinivasan, R., Mitchell, R.N., Glass, B., Hoffman, S., Rao,

S.K., Maheshwari, C., Lahiri, A., Prakash, A., McLoughlin, R., Kerner, J.K., Resnick, M.B., Montalto, M.C., Khosla, A., Wapinski, I.N., Beck, A.H., Elliott, H.L., Taylor-Weiner, A., 2021. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. Nature Communications 12, 1613. URL: https://www.nature.com/articles/s41467-021-21896-9, doi:10.1038/s41467-021-21896-9. publisher: Nature Publishing Group.

Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J.A.W.M., and the CAMELYON16 Consortium, 2017. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA 318, 2199–2210. URL: https://doi.org/10.1001/jama.2017.14585, doi:10.1001/jama.2017.14585.

Fernandez-Garza, L.E., Dominguez-Vigil, I.G., Garza-Martinez, J., Valdez-Aparicio, E.A., Barrera-Barrera, S.A., Barrera-Saldana, H.A., 2021. Personalized Medicine in Ovarian Cancer: A Perspective From Mexico. World Journal of Oncology 12, 85–92. doi:10.14740/wjon1383.

Filiot, A., Ghermi, R., Olivier, A., Jacob, P., Fidon, L., Kain, A.M., Saillard, C., Schiratti, J.B., 2023. Scaling Self-Supervised Learning for Histopathology with Masked Image Modeling. URL: https://www.medrxiv.org/content/10.1101/2023.07.21.23292757v2, doi:10.1101/2023.07.21.23292757. pages: 2023.07.21.23292757.

Fitzgerald, R.C., Antoniou, A.C., Fruk, L., Rosenfeld, N., 2022. The future of early cancer detection. Nature Medicine 28, 666–677. doi:10.1038/s41591-022-01746-x.

Gelot, C., Le-Guen, T., Ragu, S., Lopez, B.S., 2016. Double-Strand Break Repair: Homologous Recombination in Mammalian Cells, in: Kovalchuk, I., Kovalchuk, O. (Eds.), Genome Stability. Academic Press, Boston, pp. 337–351. doi:10.1016/B978-0-12-803309-8.00020-3.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum Contrast for Unsupervised Visual Representation Learning, pp. 9726–9735. doi:10.1109/CVPR42600.2020.00975. iSSN: 1063-6919.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. URL: https://ieeexplore.ieee.org/document/7780459, doi:10.1109/CVPR.2016.90. iSSN: 1063-6919.

Huang, J., Chan, W.C., Ngai, C.H., Lok, V., Zhang, L., Lucero-Prisno, D.E., Xu, W., Zheng, Z.J., Elcarte, E., Withers, M., Wong, M.C.S., 2022. Worldwide Burden, Risk Factors, and Temporal Trends of Ovarian Cancer: A Global Study. Cancers 14, 2230. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9102475/, doi:10.3390/cancers14092230.

Ilse, M., Tomczak, J., Welling, M., 2018. Attention-based Deep Multiple Instance Learning, in: Proceedings of the 35th International Conference on Machine Learning, PMLR. pp. 2127–2136. URL: https://proceedings.mlr.press/v80/ilse18a.html. iSSN: 2640-3498.

Janowczyk, A., Madabhushi, A., 2016. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. Journal of Pathology Informatics 7, 29. URL: https://www.sciencedirect.com/science/article/pii/S2153353922005478, doi:10.4103/2153-3539.186902.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet Classification with Deep Convolutional Neural Networks, in: Advances in Neural Information Processing Systems, Curran Associates, Inc. URL: https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

Lahiani, A., Klaiman, E., Grimm, O., 2018. Enabling Histopathological Annotations on Immunofluorescent Images through Virtualization of Hematoxylin and Eosin. Journal of Pathology Informatics 9, 1. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5841016/, doi:10.4103/jpi.jpi_61_17.

Lazard, T., Bataillon, G., Naylor, P., Popova, T., Bidard, F.C., Stoppa-Lyonnet, D., Stern, M.H., Decencière, E., Walter, T., Vincent-Salomon, A., 2022. Deep learning identifies morphological patterns of homologous recombination deficiency in luminal breast cancers from whole slide images. Cell Reports. Medicine 3, 100872. doi:10.1016/j.xcrm.2022.100872.

Lenz, L., Neff, C., Solimeno, C., Cogan, E.S., Abramson, V.G., Boughey, J.C., Falkson, C., Goetz, M.P., Ford, J.M., Gradishar, W.J., Jankowitz, R.C., Kaklamani, V.G., Marcom, P.K., Richardson, A.L., Storniolo, A.M., Tung, N.M., Vinayak, S., Hodgson, D.R., Lai, Z., Dearden, S., Hennessy, B.T., Mayer, E.L., Mills, G.B., Slavin, T.P., Gutin, A., Connolly, R.M., Telli, M.L., Stearns, V., Lanchbury, J.S., Timms, K.M., 2023. Identifying homologous recombination deficiency in breast cancer: genomic instability score distributions differ among breast cancer subtypes. Breast Cancer Research and Treatment 202, 191–201. doi:10.1007/s10549-023-07046-3.

Li, B., Li, Y., Eliceiri, K.W., 2021. Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14313–14323. URL: https://ieeexplore.ieee.org/document/9578683, doi:10.1109/CVPR46437.2021.01409. iSSN: 2575-7075.

Li, X., Heyer, W.D., 2008. Homologous recombination in DNA repair and DNA damage tolerance. Cell research 18, 99–113. doi:10.1038/cr.2008.1.

Litjens, G., Sánchez, C.I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., Hulsbergen van de Kaa, C., Bult, P., van Ginneken, B., van der Laak, J., 2016. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Scientific Reports 6, 26286. URL: https://www.nature.com/articles/srep26286, doi:10.1038/srep26286. publisher: Nature Publishing Group.

Liu, J.F., Konstantinopoulos, P.A., 2017. Homologous Recombination and BRCA Genes in Ovarian Cancer: Clinical Perspective of Novel Therapeutics, in: Birrer, M.J., Ceppi, L. (Eds.), Translational Advances in Gynecologic Cancers. Academic Press, Boston, pp. 111–128. doi:10.1016/B978-0-12-803741-6.00006-9.

Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G.E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P.Q., Corrado, G.S., Hipp, J.D., Peng, L., Stumpe, M.C., 2017. Detecting Cancer Metastases on Gigapixel Pathology Images. URL: http://arxiv.org/abs/1703.02442, doi:10.48550/arXiv.1703.02442. arXiv:1703.02442 [cs].

Lu, M.Y., Williamson, D.F.K., 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. URL: https://www.nature.com/articles/s41551-020-00682-w, doi:10.1038/s41551-020-00682-w. publisher: Nature Publishing Group.

Marini, N., Otálora, S., Müller, H., Atzori, M., 2021. 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. URL: https://www.sciencedirect.com/science/article/pii/S1361841521002115, doi:10.1016/j.media.2021.102165.

Miller, R.E., Leary, A., Scott, C.L., Serra, V., Lord, C.J., Bowtell, D., Chang, D.K., Garsed, D.W., Jonkers, J., Ledermann, J.A., Nik-Zainal, S., Ray-Coquard, I., Shah, S.P., Matias-Guiu, X., Swisher, E.M., Yates, L.R., 2020. ESMO recommendations on predictive biomarker testing for homologous recombination deficiency and PARP inhibitor benefit in ovarian cancer. Annals of Oncology: Official Journal of the European Society for Medical Oncology 31, 1606–1622. doi:10.1016/j.annonc.2020.08.2102.

Nero, C., Boldrini, L., Lenkowicz, J., Giudice, M.T., Piermattei, A., Inzani, F., Pasciuto, T., Minucci, A., Fagotti, A., Zannoni, G., Valentini, V., Scambia, G., 2022. Deep-Learning to Predict BRCA

Mutation and Survival from Digital H&E Slides of Epithelial Ovarian Cancer. International Journal of Molecular Sciences 23, 11326. doi:10.3390/ijms231911326.

Ngoi, N.Y.L., Tan, D.S.P., 2021. The role of homologous recombination deficiency testing in ovarian cancer and its clinical implications: do we need it? ESMO Open 6, 100144. doi:10.1016/j.esmoop.2021.100144.

Ohlén, J., Holm, A.K., 2006. Transforming desolation into consolation: being a mother with life-threatening breast cancer. Health Care for Women International 27, 18–44. doi:10.1080/07399330500377226.

Popova, T., Manié, E., Rieunier, G., Caux-Moncoutier, V., Tirapo, C., Dubois, T., Delattre, O., Sigal-Zafrani, B., Bollet, M., Longy, M., Houdayer, C., Sastre-Garau, X., Vincent-Salomon, A., Stoppa-Lyonnet, D., Stern, M.H., 2012. Ploidy and Large-Scale Genomic Instability Consistently Identify Basal-like Breast Carcinomas with BRCA1/2 Inactivation. Cancer Research 72, 5454–5462. URL: https://doi.org/10.1158/0008-5472.CAN-12-1470, doi:10.1158/0008-5472.CAN-12-1470.

Qi, Z., Redding, S., Lee, J.Y., Gibb, B., Kwon, Y., Niu, H., Gaines, W.A., Sung, P., Greene, E.C., 2015. DNA Sequence Alignment by Microhomology Sampling during Homologous Recombination. Cell 160, 856–869. doi:10.1016/j.cell.2015.01.029.

Schmauch, B., Romagnoni, A., Pronier, E., Saillard, C., Maillé, P., Calderaro, J., Kamoun, A., Sefta, M., Toldo, S., Zaslavskiy, M., Clozel, T., Moarii, M., Courtiol, P., Wainrib, G., 2020. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. Nature Communications 11, 3877. URL: https://www.nature.com/articles/s41467-020-17678-4, doi:10.1038/s41467-020-17678-4. publisher: Nature Publishing Group.

Sergeev, A., Del Balso, M., 2018. Horovod: fast and easy distributed deep learning in TensorFlow. URL: http://arxiv.org/abs/1802.05799, doi:10.48550/arXiv.1802.05799. arXiv:1802.05799 [cs, stat].

Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., zhang, y., 2021. TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification, in: Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 2136–2147. URL: https://proceedings.neurips.cc/paper/2021/hash/10c272d06794d3e5785d5e7c5356e9ff-Abstract.html.

Shi, Z., Zhao, Q., Lv, B., Qu, X., Han, X., Wang, H., Qiu, J., Hua, K., 2021. Identification of biomarkers complementary to homologous recombination deficiency for improving the clinical outcome of ovarian serous cystadenocarcinoma. Clinical and Translational Medicine 11, e399. doi:10.1002/ctm2.399.

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.

Tutt, A., Tovey, H., Cheang, M.C.U., Kernaghan, S., Kilburn, L., Gazinska, P., Owen, J., Abraham, J., Barrett, S., Barrett-Lee, P., Brown, R., Chan, S., Dowsett, M., Flanagan, J.M., Fox, L., Grigoriadis, A., Gutin, A., Harper-Wynne, C., Hatton, M.Q., Hoadley, K.A., Parikh, J., Parker, P., Perou, C.M., Roylance, R., Shah, V., Shaw, A., Smith, I.E., Timms, K.M., Wardley, A.M., Wilson, G., Gillett, C., Lanchbury, J.S., Ashworth, A., Rahman, N., Harries, M., Ellis, P., Pinder, S.E., Bliss, J.M., 2018. Carboplatin in BRCA1/2-mutated and triple-negative breast cancer BRCAness subgroups: the TNT Trial. Nature Medicine 24, 628–637. doi:10.1038/s41591-018-0009-7.

Tutt Andrew N.J., Garber Judy E., Kaufman Bella, Viale Giuseppe, Fumagalli Debora, Rastogi Priya, Gelber Richard D., de Azambuja Evandro, Fielding Anitra, Balmaña Judith, Domchek Susan M., Gelmon Karen A., Hollingsworth Simon J., Korde Larissa A., Linderholm Barbro, Bandos Hanna, Senkus Elżbieta, Suga Jennifer M., Shao Zhimin, Pippas Andrew W., Nowecki Zbigniew, Huzarski Tomasz, Ganz Patricia A., Lucas Peter C., Baker Nigel, Loibl Sibylle, McConnell Robin, Piccart Martine, Schmutzler Rita, Steger Guenther G., Costantino Joseph P., Arahmani Amal, Wolmark Norman, McFadden Eleanor, Karantza Vassiliki, Lakhani Sunil R., Yothers Greg, Campbell Christine, Geyer Charles E., 2021. Adjuvant Olaparib for Patients with BRCA1- or BRCA2-Mutated Breast Cancer. New England Journal of Medicine 384, 2394–2405. doi:10.1056/NEJMoa2105215.

Valieris, R., Amaro, L., Osório, C.A.B.d.T., Bueno, A.P., Rosales Mitrowsky, R.A., Carraro, D.M., Nunes, D.N., Dias-Neto, E., Silva, I.T.d., 2020. Deep Learning Predicts Underlying Features on Pathology Images with Therapeutic Relevance for Breast and Gastric Cancer. Cancers 12, 3687. URL: https://www.mdpi.com/2072-6694/12/12/3687, doi:10.3390/cancers12123687. number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

Veta, M., van Diest, P.J., Willems, S.M., Wang, H., Madabhushi, A., Cruz-Roa, A., Gonzalez, F., Larsen, A.B.L., Vestergaard, J.S., Dahl, A.B., Cireşan, D.C., Schmidhuber, J., Giusti, A., Gambardella, L.M., Tek, F.B., Walter, T., Wang, C.W., Kondo, S., Matuszewski, B.J., Precioso, F., Snell, V., Kittler, J., de Campos, T.E., Khan, A.M., Rajpoot, N.M., Arkoumani, E., Lacle, M.M., Viergever, M.A., Pluim, J.P.W., 2015. Assessment of algorithms for mitosis detection in breast cancer histopathology images. Medical Image Analysis 20, 237–248. URL: https://www.sciencedirect.com/science/article/pii/S1361841514001807, doi:10.1016/j.media.2014.11.010.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H., 2022. SimMIM: a Simple Framework for Masked Image Modeling, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, New Orleans, LA, USA. pp. 9643–9653. URL: https://ieeexplore.ieee.org/document/9880205/, doi:10.1109/CVPR52688.2022.00943.

Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., Xu, Y., Wei, M., Wang, W., Ma, S., Wei, F., Yang, J., Li, C., Gao, J., Rosemon, J., Bower, T., Lee, S., Weerasinghe, R., Wright, B.J., Robicsek, A., Piening, B., Bifulco, C., Wang, S., Poon, H., 2024. A whole-slide foundation model for digital pathology from real-world data. Nature , 1–8URL: https://www.nature.com/articles/s41586-024-07441-w, doi:10.1038/s41586-024-07441-w. publisher: Nature Publishing Group.

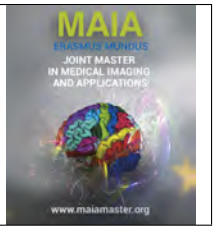
Yang, H., Chen, L., Cheng, Z., Yang, M., Wang, J., Lin, C., Wang, Y., Huang, L., Chen, Y., Peng, S., Ke, Z., Li, W., 2021. Deep learning-based six-type classifier for lung cancer and mimics from histopathological whole slide images: a retrospective study. BMC Medicine 19, 80. URL: https://doi.org/10.1186/s12916-021-01953-2, doi:10.1186/s12916-021-01953-2.

Yang, J., Nittala, M.R., Velazquez, A.E., Buddala, V., Vijayakumar, S., 2023. An Overview of the Use of Precision Population Medicine in Cancer Care: First of a Series. Cureus 15, e37889. doi:10.7759/cureus.37889.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T., 2022. iBOT: Image BERT Pre-Training with Online Tokenizer. URL: http://arxiv.org/abs/2111.07832, doi:10.48550/arXiv.2111.07832. arXiv:2111.07832 [cs].

Medical Imaging and Applications

Master Thesis, June 2024




# Scatter Correction for PET Image Reconstruction

Thitiphat Klinsuwan

*KTH, Royal Institute of Technology Stockholm*

**Abstract**

This study presents the development and implementation of a Single Scatter Simulation (SSS) algorithm for scatter correction in three-dimensional positron emission tomography (PET) imaging, carefully adhering to realistic PET imaging geometries. The algorithm, crafted in Python, is optimized with Numba and parallel processing techniques to significantly reduce the computational time associated with SSS. Moreover, to achieve superior performance, MATLAB interpolation—demonstrated to enhance computational efficiency—is invoked within Python via the MATLAB engine. The SSS algorithm's results will be integrated into iterative reconstruction techniques, specifically the Maximum Likelihood Reconstruction of Attenuation and Activity (MLAA). This research primarily aims to evaluate the impact of scatter correction utilizing the proposed Time-of-Flight (TOF) SSS algorithm and to assess the accuracy of PET imaging when TOF is combined with MLAA (TOF-MLAA). The proposed methods' performance will be initially validated using synthetic data featuring single scatter coincidences as a proof of concept. Subsequently, real clinical data, encompassing multiple scatter scenarios and acquired with the Siemens mCT Scanner, will be employed to observe the efficacy of scatter correction in a practical setting. The findings of this study are anticipated to provide valuable insights into enhancing PET imaging accuracy through the application of advanced scatter correction methodologies, thereby contributing to the field of medical imaging and improving diagnostic outcomes.

*Keywords:* Single Scatter Simulation, Scatter Correction, Iterative Reconstruction

## 1. Introduction

Positron Emission Tomography (PET) imaging has profoundly transformed medical diagnostics by offering critical insights into the physiological functions within the human body (Cherry et al., 2013). PET scanners operate on the principle of positron annihilation, detecting pairs of gamma rays emitted indirectly by a radiotracer administered to the patient. This detection occurs within a scintillator crystal, enabling the visualization and quantification of molecular-level biological processes. The selection of radiotracer is specific to the physiological event being investigated, allowing for a personalized approach to diagnosis and monitoring treatment efficacy.

The origins of Positron Emission Tomography (PET) imaging can be traced back to the 19th century, heralding the development of a transformative technology that has continually advanced. Time-of-Flight PET (TOF-PET) emerged in the 1980s and early 1990s, revolutionizing PET scanner design (Vandenberghe et al., 2020). By utilizing the variations in arrival times of emitted photons relative to their annihilation points and the detectors, TOF imaging significantly improved image resolution and contrast, thereby expanding the scope of clinical applications.

The main concept of PET imaging lies the task of reconstructing radiotracer distributions from their measured projections along lines of response (LORs). This involves solving an inverse problem to reconstruct an image from its corresponding sinogram (Natterer and Wübbeling, 2001). The forward model, represented by the attenuated Radon transform, accounts for primary photons traversing the body without attenuation. For precise reconstruction, however, it is crucial to consider scattered photons, which must be estimated during the reconstruction process.

Iterative reconstruction is a critical component in PET imaging, offering improved image quality and quantitative accuracy over traditional analytical meth-

ods. Unlike direct methods such as filtered back projection, iterative techniques refine the image through successive approximations. These algorithms, including Maximum Likelihood Expectation Maximization (MLEM) and its variants, update the image iteratively to minimize the difference between the measured and estimated projections. Maximum Likelihood Reconstruction of Attenuation and Activity (MLAA) is a notable iterative method that simultaneously reconstructs both the activity distribution and attenuation map, enhancing the accuracy of the resulting PET images. By incorporating Time-of-Flight (TOF) information, TOF-MLAA further improves the precision of image reconstruction, particularly in correcting for photon scatter.

In this study, we introduce a novel Python-based implementation of a rapid Single Scatter Simulation (SSS) algorithm. This advanced approach, referred to as scatter correction, integrates counts from Compton scattered photons into the Time-of-Flight Maximum Likelihood Reconstruction of Attenuation and Activity (TOF-MLAA) reconstruction algorithm. By leveraging this method, we aim to significantly enhance the efficiency of scatter correction processes, thereby improving the fidelity and accuracy of PET image reconstruction. This innovative technique promises to contribute substantially to the field of medical imaging, offering potential advancements in both diagnostic precision and clinical outcomes.

## 2. State of the art

During a PET scan, a radiotracer is injected into the body, which emits positrons that interact with electrons, resulting in the emission of gamma rays. These gamma rays are detected by a ring of detectors surrounding the patient. Each detector records the energy of the gamma rays at specific times, which is stored as lines of response (Beyer et al., 2000). The challenge lies in the fact that while the energy and timing of the detected gamma rays are known, the exact origin of these gamma rays within the body is not directly observable. Thus, reconstructing an image from PET data involves solving an inverse problem.

The objective of PET image reconstruction is to estimate the distribution of radiotracers within an object using measured energies captured by detectors, stored as either sinogram or list-mode data. Thus, the process of image reconstruction can be modeled as a linear inverse problem (Lewitt and Matej, 2003). The relationship between measured coincidence events $M$ during a time frame and true coincidence events $T$ is given by:

$$M = N(LT + s + r) \tag{1}$$

where $L$ and $N$ represent attenuation and normalization correction factors whiles $r$ and $s$ represent randoms and scatter contributions.

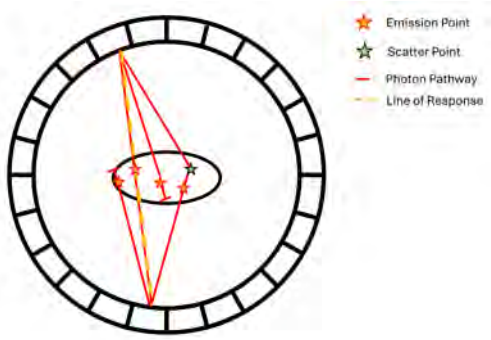



Figure 1: Illustration of different scenario involved to single LOR.

scatter contribution significantly impacts the accuracy of the emission data. When gamma rays emitted from the radiotracer within the body interact with tissues or other materials before reaching the detectors, their paths and energies are altered. This scattering leads to mispositioning and inaccurate energy readings, which can result in erroneous data being recorded by the detectors. The presence of scattered photons increases background noise and reduces the signal-to-noise ratio, making it challenging to accurately reconstruct the spatial distribution of the radiotracer. Correcting for scatter is essential to enhance image quality and ensure precise quantitative analysis of the metabolic activity being studied (Cherry et al., 2013).

Consequently, the scatter correction is crucial for enhancing the contrast and accuracy of reconstructed PET images, significantly improving the quantification of activity within the body (Barney et al., 1991). Early research on scatter correction strategies focused on modeling and compensating for scatter in three-dimensional PET imaging, incorporating foundational concepts to mitigate scatter effects (Barney et al., 1991; Buvat et al., 1994; Zaidi and Koral, 2004). Subsequent studies extended these strategies by modeling multiple scatter events through Gaussian smoothing applied to a simulated single scatter sinogram, providing a more refined approach to scatter correction (Goggin and Ollinger, 1994; Ollinger and Johns, 1993). Additionally, the use of Monte Carlo simulations to estimate scatter distributions has been proposed for clinical applications, despite the associated increase in computational demands (Holdsworth et al., 2003; Levin et al., 1995).

The primary focus of this study is the implementation of a single scatter modeling algorithm that accounts for multiple scatter effects by scaling the estimated single scatter sinogram to match the measured data. This method has been highlighted in various sources for its theoretical advantage of reduced computational time while maintaining effective estimation accuracy (Ollinger and Johns, 1993; Panin, 2012; Thielemans et al., 2007; Watson et al., 1996). Notably, this approach has been recognized as desirable approach due

to its computational efficiency, as detailed in (Watson, 1999). By leveraging this algorithm, the study aims to provide a robust and efficient solution for scatter correction in PET imaging, ultimately enhancing the overall quality and reliability of the reconstructed images.

The effectiveness of scatter-compensation techniques is fundamentally reliant on the precision of scatter estimation. An extensive review by (Zaidi and Koral, 2004) categorizes scatter correction approaches into five distinct groups, highlighting the diversity and complexity of existing methodologies. For this study, we will predominantly focus on statistical iterative reconstruction techniques, which are well-documented in the literature and form the basis of many contemporary algorithms (Hutton et al., 2006). These methods are favored for their ability to integrate sophisticated models of scatter and attenuation, thereby enhancing image quality and quantitative accuracy. By leveraging these advanced iterative techniques, our research aims to achieve more accurate scatter correction, ultimately contributing to the development of more reliable and precise PET imaging modalities.

## 3. Material and methods

### *3.1. Background*

Positron Emission Tomography (PET) acquisition is fundamentally an inverse problem. In PET, the goal is to reconstruct an image representing the distribution of a radiotracer within the body from the detected gamma rays. This problem is considered inverse because we do not know the exact emitter points of the photons; instead, we work backward from the detected signals to infer the source distribution (Cherry et al., 2013).

In PET imaging, data acquisition can be performed in different modes, one of which is list mode. List mode data acquisition involves recording each detected event individually, storing the exact time of detection, the position of the detectors involved, and the energy of the detected photons. This mode provides the most detailed information about each event, allowing for flexible post-processing and reconstruction methods.

The Line of Response (LOR) is a critical concept in PET imaging. When a positron emitted by the radiotracer undergoes annihilation with an electron, two gamma photons are emitted simultaneously in approximately opposite directions. These photons are detected by the PET scanner, and the line connecting the two detection points is referred to as the LOR. The LOR represents the path along which the annihilation event occurred. By collecting multiple LORs from different angles, it is possible to reconstruct the spatial distribution of the radiotracer within the body.

The mathematical foundation for summing along photon paths is the Radon transform. The Radon transform is a mathematical integral transform that converts

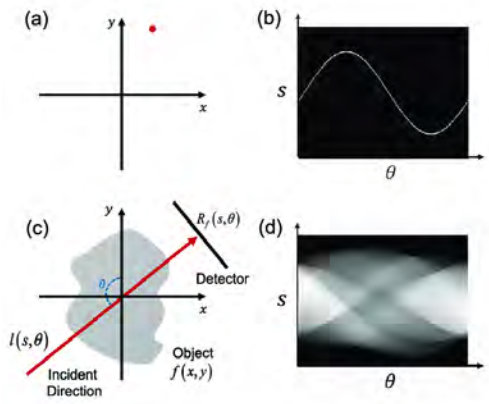



Figure 2: Illustration of the Radon transform: (a) A point source in Cartesian coordinates; (b) The Radon transform of a point source, represented as a sinusoidal wave; (c) Projection of an object $f(x, y)$ along the line $l(s, \theta)$ at angle $\theta$; (d) The Radon transform of the object, showing the projection data $R_f(s, \theta)$ as a function of $s$ and $\theta$. The image taken from (Zuo et al., 2020).

a spatial domain function into a set of projections. It is fundamental in various imaging techniques, including computed tomography (CT) and positron emission tomography (PET).

For a function $f(x, y)$ in two dimensions, the Radon transform $Af$ is defined as the integral of $f$ over lines. Mathematically, it is expressed as:

$$Af(\theta, t) = \int_{-\infty}^{\infty} f(x\cos\theta + y\sin\theta = t)\, ds, \qquad (2)$$

where $\theta$ represents the angle of the line and $t$ is the perpendicular distance from the origin to the line. This transform essentially gathers all line integrals of $f$ at different angles and distances, generating a set of projections (Deans, 2007).

In three dimensions, the Radon transform extends to the integration over planes. For a function $f(x, y, z)$, the 3D Radon transform $Af$ is given by:

$$Af(\theta, \phi, t) = \int_{\mathbb{R}^2} f(x\cos\theta\sin\phi + y\sin\theta\sin\phi + z\cos\phi)\, d\sigma, \qquad (3)$$

where $(\theta, \phi)$ defines the plane orientation and $t$ is the distance from the origin to the plane.

Forward projection in 3D PET imaging can be modeled using the Radon transform. Given a radiotracer distribution $f(x, y, z)$, the projection data $p(\theta, \phi, t)$ along a line can be computed as:

$$p(\theta, \phi, t) = \int_{-\infty}^{\infty} f(x\cos\theta\sin\phi + y\sin\theta\sin\phi + z\cos\phi)\, dz. \qquad (4)$$

This integral represents the total activity along the specified plane. The PET detectors collect these plane integrals over various angles and distances, forming a sinogram used for image reconstruction. The forward projection process can be computationally intensive due

Table 1: Annotation Table

| Symbol | Description | Symbol | Description |
|---|---|---|---|
| $A$ | Radon transform | $L$ | Attenuation correction factor |
| $N$ | Normalization factor | $s$ | Scatter prompts |
| $r$ | Random prompts | $\triangle s$ | Spatial offset |
| $\sigma$ | Standard deviation, Geometrical cross section | $\mu$ | Mean |
| $\mu_l$ | Linear attenuation | $\mu_m$ | Mass attenuation |
| $\rho$ | Density | $h$ | Planck constant |
| $c$ | Speed of light | $v$ | Velocity |
| $\lambda$ | Wavelength | $\lambda_m$ | Attenuation map |
| $\rho_m$ | Attenuation map | $M$ | Emission prompts |
| $\overline{M}$ | Estimated Emission prompts | $T$ | True prompts |

to the large datasets involved in 3D PET imaging, often requiring efficient algorithms and approximations to manage the computational load (Kak and Slaney, 2001).

Time-of-Flight (TOF) technology in Positron Emission Tomography (PET) provides a significant advancement in image reconstruction by estimating the origin of positron annihilation events. TOF-PET determines the position of the event based on the difference in arrival times of the emitted photons at the detectors. The technique assumes the annihilation photons travel at the speed of light ($c$), and any difference in detection time ($\Delta t$) is due to the difference in path lengths from the event to each detector (Budinger, 1983).

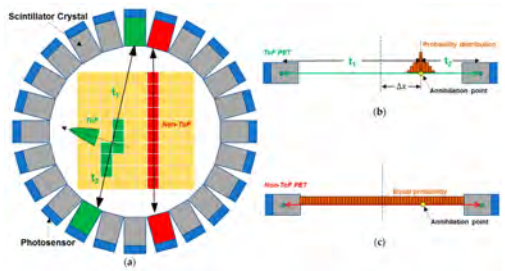



Figure 3: Concept of time-of-flight positron emission tomography (ToF PET): (a) Detector ring detecting gamma photon pairs with (green) and without (red) ToF; (b) Probability distribution of the annihilation position along the line of response (LoR) in ToF PET; (c) Equal probability of annihilation position along the LoR in non-ToF PET. Source: (Jiang et al., 2019).

The detection efficiency for TOF measurements can be approximated by a quasi-Gaussian function, as is common in the literature, despite the discrete nature of the acquired data. This function is given by:

$$\varepsilon_t(\Delta s) = e^{-\frac{(\Delta s - t\sigma)^2}{2\sigma^2}} \Bigg/ \sum_{t'} e^{-\frac{(\Delta s - t'\sigma)^2}{2\sigma^2}}, \tag{5}$$

where $\Delta s$ represents the spatial offset from the midpoint between the two detectors to the annihilation event, $t$ indicates the TOF bin width, and $\sigma$ symbolizes the system's timing resolution. This efficiency function effectively convolves the intrinsic TOF resolution with the square function of a TOF bin, modeling the probability that a detected emission event with an offset of $\Delta s$ will be recorded in the $t$-th TOF bin.

The full width at half maximum (FWHM) of the TOF resolution is related to the timing resolution $\sigma$ by:

$$\text{FWHM} = 2\sqrt{2\ln(2)}\sigma, \tag{6}$$

which reflects the precision of event localization. The improved localization provided by TOF contributes to the enhanced signal-to-noise ratio and the superior image quality in TOF-PET.

Highlighting that a diminished FWHM enhances energy resolution, allowing for finer discrimination between energy levels. Energy resolution ($R$) itself is then articulated in relation to FWHM and the mean energy ($\mu$) as:

$$R = \frac{\text{FWHM}}{\mu} \tag{7}$$

With lower $R$ values indicating superior energy resolution, this measure underscores the detector's ability to precisely gauge energy values, setting the stage for groundbreaking experimental accuracy.

Turning our focus to detector efficiency, this aspect evaluates the likelihood of accurately detecting a photon of a specified energy, depicted through the cumulative distribution function (CDF):

$$\Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right] \tag{8}$$

The CDF, $\Phi$, thus maps the cumulative probability for detected energy to be at or below $x$, gradually approaching unity, which signifies impeccable detection efficiency. The integral role of the error function (erf) in the CDF is outlined as:

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt \tag{9}$$

The intricate dance between PDF and CDF, influenced by $\sigma$ and FWHM, forms the analytical backbone

for evaluating energy resolution and detector efficiency. This nuanced statistical interplay not only informs the design and optimization of detectors but also empowers scientists to navigate the complex terrain of experimental physics with enhanced precision and understanding. By harnessing these mathematical constructs, the quest for accuracy in energy measurements propels forward, shaping the future of experimental endeavors.

For a single line of response in a PET scanner, there are several contributions that reach the crystal or detector. These contributions occur due to the activity of photons passing through various materials. Although photons generally travel in a straight line, they can interact with atomic masses along their path. There are several possible ways for these interactions to occur, each with its own consequences.

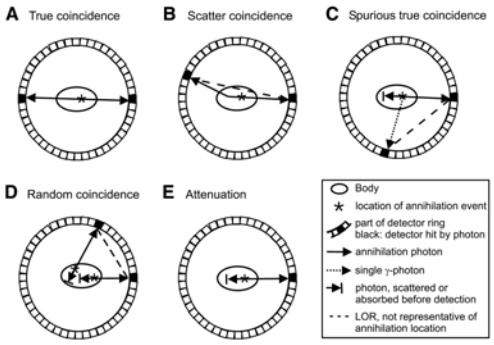



Figure 4: Different types of coincidence events in PET imaging. The figure is taken from (Verel et al., 2005)

Photon activity, specifically each photon's energy, affects the cross-section with respect to power versus atomic number. In nuclear medicine, the interactions between photons and matter are crucial for various imaging and therapeutic applications. Photons generally travel in straight lines but can interact with atomic masses along their path, resulting in several possible interactions. While there are nine potential interactions between photons and matter, we focus on three significant interactions that are particularly relevant to PET (Positron Emission Tomography). These interactions will be discussed in detail, covering both their physical mechanisms and mathematical aspects.

The interaction cross-section of photons with atoms is highly dependent on the photon's energy and the atomic number ($Z$) of the material. Higher atomic numbers increase the probability of certain types of interactions, such as photoelectric absorption. The cross-section for photoelectric absorption ($\sigma_{\text{photo}}$) is approximately proportional to $Z^3$ and inversely proportional to the photon energy ($E$) raised to the power of three:

$$\sigma_{\text{photo}} \propto \frac{Z^3}{E^3}$$

In contrast, Compton scattering, which involves the photon's interaction with an electron, is less dependent on the atomic number and more on the electron density of the material.

The linear attenuation coefficient ($\mu_l$) characterizes the extent to which a material attenuates photons. This coefficient combines the contributions from all possible interaction processes: photoelectric absorption, Compton scattering, and pair production. The linear attenuation coefficient can be expressed as:

$$\mu_l = \mu_{\text{photo}} + \mu_{\text{Compton}} + \mu_{\text{pair}}$$

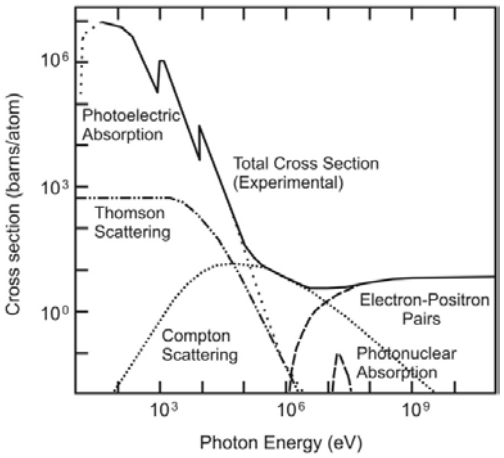



Figure 5: Photon-energy dependent cross sections. Cross sections of the photoelectric absorption, Thomson scattering, Compton scattering, electron-positron pairs and photonuclear absorption for Cu as a function of energy. The figure is taken from (Hermanns, 2013).

Additionally, the mass attenuation coefficient ($\mu_m$), which normalizes the linear attenuation coefficient by the density ($\rho$) of the material, provides a measure of how a specific material attenuates photons regardless of its density:

$$\mu_m = \frac{\mu_l}{\rho}$$

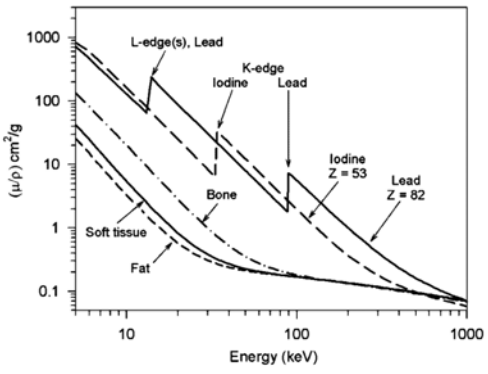



Figure 6: Mass attenuation coefficient. The image is taken from (Seibert and Boone, 2005)

- **Photoelectric Absorption**: This process dominates at lower photon energies, where a photon is completely absorbed, transferring its energy to an electron, which is then ejected from the atom. The likelihood of photoelectric absorption decreases with increasing photon energy and is highly dependent on the atomic number of the absorbing material.

- **Thomson Scattering**: Thomson scattering occurs when a photon interacts with a free electron, causing the photon to be deflected without a change in its energy. However, in the photon energy range used in PET, which is typically around 511 keV, the contribution from Thomson scattering is minimal. This is because the scattering cross-section decreases significantly at these higher energies, making other interactions, such as Compton scattering, more predominant.

- **Compton Scattering**: Predominant at intermediate energies, Compton scattering involves the photon transferring part of its energy to an electron and being deflected in a different direction. This scatter of photons is a significant factor in PET imaging, contributing to image degradation unless corrected for.

Compton scattering describes the interaction between a photon and a loosely bound outer-shell orbital electron of an atom. In this process, the incident photon, with energy much greater than the binding energy of the electron, effectively behaves as if it collides with a free electron. Unlike in the photoelectric effect, the photon does not vanish; instead, it undergoes deflection through a scattering angle ($\theta$). During Compton scattering, a portion of the photon's energy is imparted to the recoil electron, resulting in a decrease in the photon's energy.

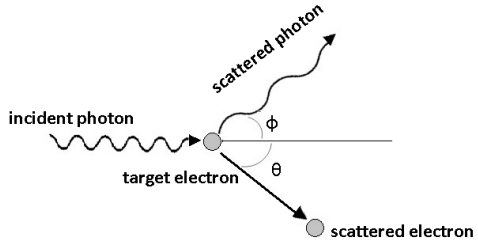



Figure 7: Compton scattering

The relationship between the energy of the scattered photon ($E'$) and the scattering angle is governed by the principles of energy and momentum conservation. Specifically, the energy of the scattered photon is given by the equation:

$$E' = \frac{E}{1 + \left(\frac{E}{m_e c^2}\right)(1 - \cos(\theta))}$$

Here, $E$ represents the energy of the incident photon in MeV. It is important to note that the energy transferred during Compton scattering is independent of the properties of the absorbing material, such as density or atomic number. Additionally, Compton scattering strictly involves interactions between photons and electrons, with no dependence on other characteristics of the material.

The Klein-Nishina formula (Klein and Nishina, 1928) provides the differential cross-section for photons scattered by a single free electron. This formula, derived within the framework of quantum electrodynamics, represents one of the earliest successful applications of the Dirac equation. It describes the scattering of photons by electrons, accounting for both Thomson scattering at low photon energies and Compton scattering at high photon energies.

$$\frac{d\sigma}{d\Omega} = \frac{1}{2} r_e^2 \left(\frac{\lambda}{\lambda'}\right)^2 \left(\frac{\lambda}{\lambda'} + \frac{\lambda'}{\lambda} - \sin^2(\theta)\right)$$

The angular dependent photon wavelength (or energy, or frequency) ratio is

$$\frac{\lambda}{\lambda'} = \frac{E'_\gamma}{E_\gamma} = \frac{\omega'}{\omega} = \frac{1}{1 + \frac{h\nu}{m_e c^2}(1 - \cos\theta)}$$

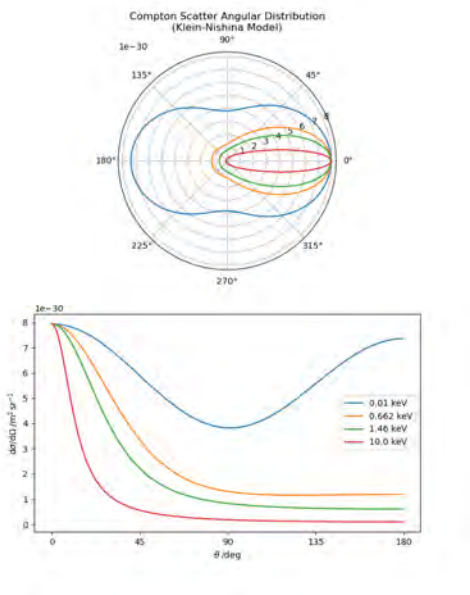



Figure 8: The Klein–Nishina predictions of photon scattering. This image is sourced from (Hill, 2019).

The formula elucidates how the total cross-section and the expected deflection angle of scattered photons change with increasing photon energy. Notably, it

demonstrates that at higher photon energies, the cross-section decreases, indicating a reduced likelihood of interaction between photons and electrons as shown in Figure 6. This insight is crucial for understanding various phenomena in particle physics, astrophysics, and medical imaging, where the scattering of photons plays a significant role.

In TOF PET tomography (Watson, 2005), the arrival-time difference $\Delta t$ is related to the spatial offset $\Delta s$ by $\Delta t = 2\Delta s/c$. This relationship is integrated into the SSS model by a detection efficiency function $\varepsilon_t(\Delta s)$, the contribution of the scatter signal to the detected events is modeled through a set of integrals accounting for various physical phenomena:

$$S_{AB} = \int_{V_s} \frac{\sigma_A \sigma_B}{4\pi R_{AS}^2 R_{BS}^2} \frac{\mu_l}{\sigma_c} \frac{d\sigma_c}{d\Omega} [I_A + I_B] \, dV_s \qquad (10)$$

$$I_A = e^{-(\int_A^S \rho_m ds + \int_B^S \rho'_m ds)} \int_A^S \varepsilon_t(R_{BS} - R_{AS} + 2s) \lambda_m(s) ds$$

$$I_B = e^{-(\int_B^S \rho_m ds + \int_A^S \rho'_m ds)} \int_B^S \varepsilon_t(R_{BS} - R_{AS} - 2s) \lambda_m(s) ds$$

where each term is defined as follows:

- $S_{AB}$: Scatter signal detected by detectors A and B from a volume element $dV_s$.
- $\frac{\sigma_A \sigma_B}{4\pi R_{AS}^2 R_{BS}^2}$: Geometrical efficiency, indicating the effectiveness with which the system detects photons that have scattered once at the point S and are then captured by the detector pair.
- $\frac{\mu_l}{\sigma_c}$: This term describes the normalized linear attenuation coefficient $\mu$, where $\sigma_c$ is the total Compton scattering cross-section. The linear attenuation coefficient $\mu_l$ quantifies how much a material can attenuate the intensity of the radiation passing through it, typically measured in $\text{cm}^{-1}$. Normalizing it by $\sigma_c$ used to express the attenuation relative to the probability of Compton scattering events, effectively scaling the attenuation by the scattering interactions in the medium.
- $I_A$ and $I_B$: Integrals representing the attenuation of unscattered photons along the paths to detectors A and B, modified by the TOF detection efficiency function $\varepsilon_t(\Delta s)$.

Nonetheless, the single scatter simulation will be corrupted by random coincidences, as seen in tailing outside the attenuation mask of emission data, as explained in (Watson et al., 2004). The example of this artifact shown in 9.

After acquiring the scatter sinogram from the simulation using the activity map, we compute the scaling factor using a linear fitting with $Y = mx$, where $x$ represents the scatter sinogram with the attenuation mask and $Y$ represents the tail sinogram computed from the difference of total prompts and random prompts. An example of this process is shown in Figure 23.

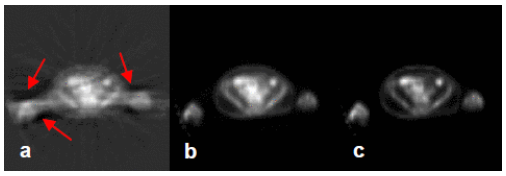



Figure 9: Example of estimated images used for scatter simulation with the random coincidence effect. The images are sourced from (Watson et al., 2004). (a) DIFT, (b) OSEM for the first scatter iteration, and (c) second OSEM iteration. Note the negative regions in the DIFT image (indicated by arrows), contributed to by scatter and patient arm motion.

### *3.2. PET Reconstruction*

#### *3.2.1. Maximum-Likelihood Expectation Maximization (MLEM)*

The reconstruction process in PET imaging is essential for accurately updating the activity map, which represents the distribution of the radiotracer within the body. The MLEM algorithm (Lange and Carson, 1984; Shepp and Vardi, 1982) is a statistical method used to iteratively refine this activity map $\lambda^{(k+1)}$ by maximizing the likelihood of the measured data given the current estimate of the activity distribution.

The basic MLEM update equation is given by:

$$\lambda_m^{k+1} = \frac{\lambda_m^k}{A^T 1} \cdot A^T \left( \frac{M}{A\lambda_m^k + b} \right) \qquad (11)$$

where:

- $\lambda_m^k$ is the current estimate of the activity map.
- $A$ is the forward projection operator, which models the PET scanner's response to the activity distribution.
- $b$ represents the background noise, modeled as ordinary Poisson noise.
- $M$ is the measured data, i.e., the detected PET events.

In real PET imaging scenarios, several additional factors need to be incorporated to improve the accuracy of the reconstruction:

- Normalization correction factor ($N$): Accounts for variations in detector efficiencies and geometrical misalignment.
- Attenuation factor ($L$): Compensates for the attenuation of photons as they pass through the body.

- Scatter prompts ($s$): Estimates the contribution of scattered photons to the detected signal.
- Random prompts from delayed events ($r$): Corrects for random coincidences in the detected signal.

Considering these parameters, the iterative reconstruction model is updated as follows:

$$\lambda_m^{k+1} = \frac{\lambda_m^k}{(A^T L^T N^T 1)} \cdot A^T L^T N^T \left( \frac{M}{NLA\lambda_m^k + s + r} \right) \quad (12)$$

Here, the additional factors $L$, $N$, $s$, and $r$ are incorporated into the forward and backward projection operations, ensuring a more accurate reconstruction of the activity map.

### 3.2.2. Maximum-Likelihood Transmission Reconstruction (MLTR)

The MLTR algorithm (Manglos et al., 1995) is employed for attenuation correction by utilizing the separation of transmission and emission data, particularly with the aid of Time-of-Flight (TOF) information. The attenuation correction is vital for accurate quantitative PET imaging, as it accounts for the loss of photon pairs due to absorption in the body tissues.

The process involves updating the attenuation map $\sigma$, which represents the distribution of attenuation coefficients within the body. This update is based on the current estimates of the activity map $\lambda$, attenuation map $\sigma$, forward projection $A$, normalization term $N$, measured data $m$, random estimate $r$, scatter estimation $s$, and sensitivity term $S$.

The updated term $\psi$ is calculated as:

$$\psi = N \cdot e^{-A(\rho_m^{(k)})} \cdot A(\lambda_m)$$

The update equation for the attenuation map $\sigma$ in MLTR is given by:

$$\rho_m^{k+1} = \rho_m^k + \frac{A^T \left( \frac{\psi}{\psi + r + s - m} \right)}{A^T \left( \frac{\psi^2}{\psi + r + s} \right)} \cdot S$$

where:

- $\rho_m^k$ is the current estimate of the attenuation map.
- $\psi$ represents the expected projections given the current estimates.
- $S$ is the sensitivity term, which accounts for the detector sensitivity variations.

This iterative update ensures that the attenuation map accurately reflects the true distribution of attenuation coefficients, thereby improving the overall quality of the PET reconstruction.

### 3.2.3. Joint Reconstruction of Activity and Attenuation in Time-of-Flight PET (MLAA)

The MLAA algorithm (Benoit et al., 2016) combines the iterative reconstruction methods of MLEM and MLTR to simultaneously update both the activity map $\lambda$ and the attenuation map $\sigma$. This joint reconstruction approach leverages the additional information provided by TOF PET, which enhances the quantitative accuracy and spatial resolution of the reconstructed images.

**Algorithm 1** The joint 3-algorithm with updates in simplified notation

1: $\lambda_m \leftarrow \lambda_m^{\text{init}}, \rho_m \leftarrow \rho_m^{\text{init}}$ ▷ Initialization
2: **for** each iteration **do**
3: $\lambda_m \leftarrow \lambda_m \cdot \frac{A^T L^T N^T (m/\tilde{m})}{A^T L^T N^T 1}$ ▷ Sub-iteration 1: MLEM
4: $\rho_m \leftarrow \rho_m + \frac{A^T (\psi/\tilde{m} \cdot (\tilde{m} - m))}{A^T (\psi^2/\tilde{m} \cdot S)}$ ▷ Sub-iteration 2: MLTR
5: $s \leftarrow SSS(\lambda_m, \rho_m)$ ▷ Sub-iteration 3: SSS
6: **end for**

The MLAA update equations for the activity map and attenuation map are as follows:

$$\lambda^{(k+1)} = \frac{\lambda^{(k)}}{(A^T L^T N^T 1)} \cdot A^T L^T N^T \left( \frac{m}{NLA\lambda^{(k)} + s + r} \right)$$

$$\sigma^{(k+1)} = \sigma^{(k)} + \frac{A^T \left( \frac{\psi}{\psi + r + s - m} \right)}{A^T \left( \frac{\psi^2}{\psi + r + s} \right)} \cdot S$$

In this framework, $\lambda_m$ and $\rho_m$ are iteratively updated in an alternating fashion. The activity map update ($\lambda_m$) is performed using the MLEM approach, incorporating the attenuation correction and other factors. The attenuation map update ($\rho_m$) is done using the MLTR approach, ensuring that both maps are simultaneously refined. This joint reconstruction leads to improved image quality and quantitative accuracy, making it particularly useful for advanced PET imaging applications.

## 3.3. Evaluation Metrics

This section delineates the evaluation metrics employed to assess the performance of our reconstruction method, specifically the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index (SSIM). These metrics are applied in both the image space and the sinogram space to provide a comprehensive evaluation.

### 3.3.1. Evaluation in Image Space

In the image space, the evaluation is conducted by computing the Signal-to-Noise Ratio (SNR) from the reconstructed image alone, without the need for a reference image or ground truth. This metric evaluates the consistency of the reconstructed image and estimates

noise by comparing the differences between the reconstructed image and its smoothed version using a Gaussian filter.

The Signal-to-Noise Ratio (SNR) can be computed by comparing the mean of the signal to the standard deviation of the noise. The steps involved in this computation are as follows:

- Noise Estimation: Apply a 3D smoothing filter (e.g., Gaussian filter) to the original image to create a smoothed version. The difference between the original and smoothed images is considered the noise.

$$S(i, j, k) = (I * G)(i, j, k) \tag{13}$$

where $G$ is the 3D Gaussian filter, and $*$ denotes the convolution operation.

$$N(i, j, k) = I(i, j, k) - S(i, j, k) \tag{14}$$

- Calculate the Mean of the Original Image: Compute the mean of the original 3D image.

$$\mu_{\text{signal}} = \frac{1}{P \cdot Q \cdot R} \sum_{i=0}^{P-1} \sum_{j=0}^{Q-1} \sum_{k=0}^{R-1} I(i, j, k) \tag{15}$$

- Calculate the Standard Deviation of the Noise: Compute the standard deviation of the noise estimated from the original and smoothed images.

$$\sigma_{\text{noise}} = \sqrt{\frac{1}{P \cdot Q \cdot R} \sum_{i=0}^{P-1} \sum_{j=0}^{Q-1} \sum_{k=0}^{R-1} [N(i, j, k)]^2} \tag{16}$$

- Compute SNR: Use the ratio of the mean of the original image to the standard deviation of the noise.

$$\text{SNR} = \frac{\mu_{\text{signal}}}{\sigma_{\text{noise}}} \tag{17}$$

By following these steps and using the provided equations, the SNR for a 3D image can be estimated without a reference image or ground truth, thereby providing a measure of the image's consistency. This method allows for a robust evaluation of image quality in the absence of an external standard, ensuring that the intrinsic properties of the image are adequately assessed.

*3.3.2. Evaluation in Sinogram Space*

In the sinogram space, evaluation involves comparing the expected sinogram with the measured sinogram using both PSNR and SSIM metrics. This method offers a robust assessment of image quality and structural similarity in the transformed domain.

Peak Signal-to-Noise Ratio (PSNR) (Sheikh et al., 2006) is used to evaluated the estimated sinogram with the measure data in sinogram space. The Mean Squared Error (MSE) is computed between the expected and measured sinograms as follows:

$$\text{PSNR} = 10 \log_{10} \left( \frac{\text{MAX}^2}{\text{MSE}} \right) \tag{18}$$

$$\text{MSE} = \frac{1}{P \cdot Q \cdot R} \sum_{i=0}^{P-1} \sum_{j=0}^{Q-1} \sum_{k=0}^{R-1} [I(i, j, k) - K(i, j, k)]^2 \tag{19}$$

- MAX: The maximum possible pixel value of the image. For example, for an 8-bit image, this value is 255.
- MSE: Mean Squared Error between the original and reconstructed 3D images.
- $P$: The number of rows (height) in the 3D image.
- $Q$: The number of columns (width) in the 3D image.
- $R$: The depth (number of slices) in the 3D image.
- $I(i, j, k)$: The pixel value at position $(i, j, k)$ in the original 3D image.
- $K(i, j, k)$: The pixel value at position $(i, j, k)$ in the reconstructed 3D image.

Structural Similarity Index (SSIM), (Wang and Bovik, 2002) is a perceptual metric that evaluates the similarity between two images by considering changes in structural information, luminance, and contrast. For sinograms, SSIM is calculated as follows:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{20}$$

where $x$ and $y$ are the expected and measured sinograms, $\mu_x$ and $\mu_y$ represent their mean values, $\sigma_x^2$ and $\sigma_y^2$ are their variances, $\sigma_{xy}$ is their covariance, and $C_1$ and $C_2$ are constants to avoid division by zero. The SSIM index ranges from -1 to 1, with 1 indicating perfect similarity. SSIM is considered more consistent with human visual perception than PSNR.

To ensure a thorough evaluation, the PSNR and SSIM scores are computed for each bin of the sinogram, and the final reported values are averages over all bins. This approach ensures a robust assessment of image quality and structural similarity across different data representations, thus enhancing the reliability of the reconstruction method.

In summary, PSNR provides a measure of absolute error, while SSIM offers a perceptual evaluation of image quality in the sinogram space. Utilizing both metrics allows for a comprehensive assessment of the quantitative and perceptual aspects of image reconstruction, ensuring the efficacy of the proposed method.

### *3.4. Computational Resources*

The experiments were validated using MATLAB R2024a (Update 3) on a Linux-based system with an Intel(R) Xeon(R) Platinum 8375C CPU @ 2.90GHz. The system architecture was x86_64, supporting 32-bit and 64-bit operations with 46-bit physical and 48-bit virtual address sizes. It had 16 CPUs (8 cores with 2 threads per core). The openSSS, an open-source implementation of scatter estimation for 3D TOF-PET based on the TOF-aware Single Scatter Simulation (SSS), was tested and timed using this MATLAB setup, ensuring sufficient computational capacity and data handling for accurate experimental results.

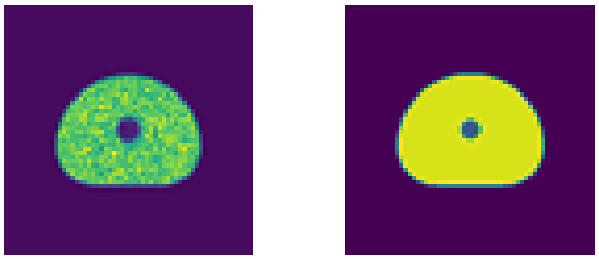
On the other hand, for python experiment with larger clinical data. The experiments were conducted on a laboratory server equipped with a 48-core Intel Xeon E5-2690 v3 processor (2.60 GHz base clock, 3.50 GHz turbo) supporting both 32-bit and 64-bit instruction sets. This high-performance computing environment features a 60 MiB shared L3 cache and hardware virtualization capabilities (VT-x).

## 4. Results

### *4.1. Single Scatter Simulation*

The development of the single scatter simulation was successfully implemented using Python, with the results meticulously benchmarked against estimations obtained from the MATLAB version available online (Santo et al., 2023). Furthermore, comparisons with the Monte Carlo simulation are currently underway to ensure comprehensive validation, in collaboration with the developer of openSSS as a partner.

To accurately determine the sinogram, it is imperative to establish both the activity map and attenuation map, alongside the scanner geometry. In order to minimize computational time, a subset of detectors and rings was uniformly selected from the complete set available in the scanner. For the Toyscanner, which comprises 320 detectors and 8 rings, the presented results are computed using 80 sample detectors and 3 rings. The activity map and attenuation map are depicted in Figure 10a and 10b, respectively, providing a visual representation of the spatial distribution of radioactive tracers and the attenuation properties of the scanned object.



(a) Activity Map (b) Attenuation Map

Figure 10: Activity and Attenuation Maps

Figure 11 illustrates the scatter distribution contribution from the sample ring and detector, specifically aimed at reducing computational time. This figure highlights the scatter events detected within a specific subset of the scanner, providing a focused view of scatter behavior. Additionally, the interpolated sinogram, or estimated scatter distribution for all rings and detectors, was computed using multi-linear interpolation, as shown in Figure 12. This approach allows for the estimation of scatter distribution across the entire scanner, enhancing the accuracy of the simulation.

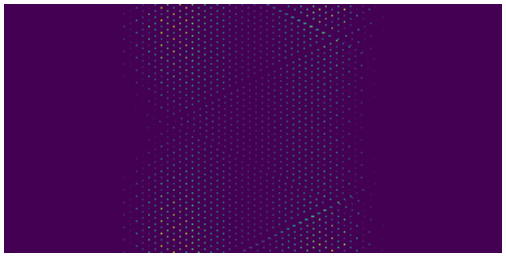


Figure 11: Single Scatter Simulation with sample detectors and rings

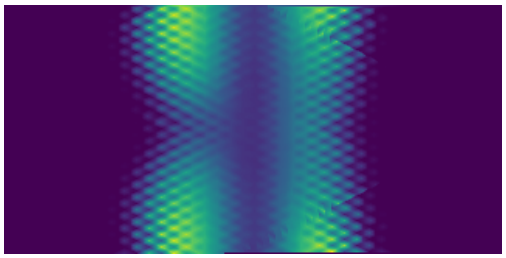


Figure 12: Interpolated Single Scatter Simulation

The contribution of Time-of-Flight Single Scatter Simulation (TOF-SSS) is depicted in Figure 12. The results indicate that the contributions from edges of FOV are smaller than those from center of FOV, corresponding to the normal distribution described in Equation 5. This finding aligns with theoretical expectations and demonstrates the efficacy of TOF-SSS in capturing scatter events within specific temporal windows.

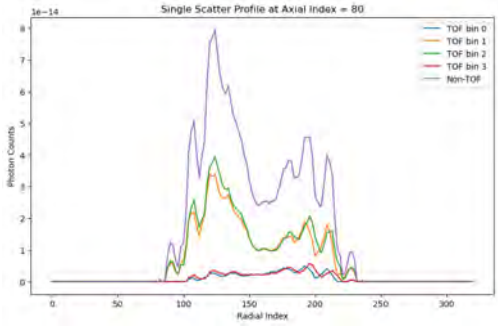



Figure 13: Single Scatter Profile at Axial Index = 80

To further validate the performance of the Python im-

plementation of Single Scatter Simulation (SSS), we evaluated the computational times for SSS using MATLAB, standard Python (without utilizing Numba and parallel processing), and optimized Python (utilizing Numba and parallel processing). These evaluations are presented in Figure 14. The results indicate that for the SSS part computed from the sample ring and detector, the optimized Python version is nearly 25 times faster than the MATLAB version, while the standard Python version is 8 times slower. However, the interpolation part in MATLAB, using the interp function, is 5 times faster than the Python version. This significant improvement in computational efficiency underscores the benefits of using optimized Python implementations for large-scale simulations.

This SSS algorithm will be used to compute the scatter sinogram for the reconstruction process, with the results to be presented in the subsequent section.

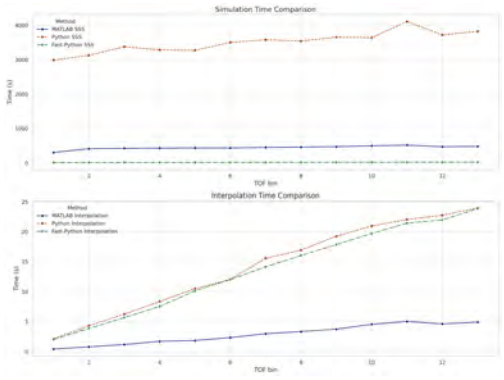



Figure 14: Computational Time of Single Scatter Simulation

### 4.2. Emission Sinogram Simulation

To validate the concept of scatter correction, we present the generation of the emission sinogram using generated scatter sinograms and forward projection of the activity map. This process follows Equation 1, ensuring a systematic approach to simulating emission data. The attenuation term is generated using the following equation:

$$L = e^{-A(\rho_m)} \tag{21}$$

Scatter sinograms are generated using the activity and attenuation map phantom with the same Toy Scanner Geometry and 5 TOF bins. A random sinogram is generated using a Poisson distribution with $\lambda = 0.5$. In the final step, Poisson noise is added to the generated emission sinogram to mimic real-world data acquisition scenarios. The result of this generation process is shown in Figure 15, and its corresponding profile is illustrated in Figure 16. These figures provide a comprehensive view of the simulated emission data, showcasing the effects of scatter and random events.

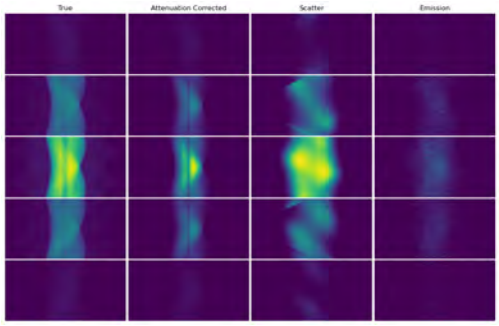



Figure 15: Synthetic Emission Sinogram where each row represent each TOF bin

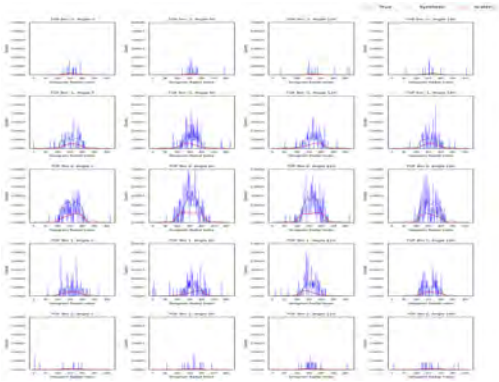


Figure 16: Emission and scatter sinogram profiles at different Bin and Angle, each row represents TOF bin and column represents angle 0, 60, 120, 180, respectively

### 4.3. Scatter Correction

To provide a comprehensive understanding of the scatter correction process, we detail each argument of the Maximum Likelihood Expectation Maximization (MLEM) algorithm. Initially, the attenuation sinogram is computed and used for forward projection to calculate the denoising term, as described in Equation 12, but specifically for the case without scatter correction.

Figure 17: Attenuation Sinogram $L = e^{-A(\rho_m)}$

After obtaining the mask by computing the attenuation sinogram, we compute the sensitivity or denoising term according to the divisor term $L^T A^T 1$. This term plays a crucial role in normalizing the reconstructed images, ensuring accurate representation of tracer distribution.

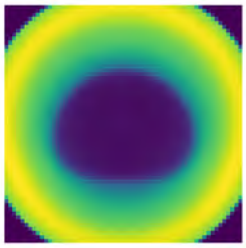


Figure 18: Sensitivity of MLEM

The initial guess for the iterative reconstruction is given by $LA(\lambda_m) + r$, providing a starting point for the MLEM algorithm.

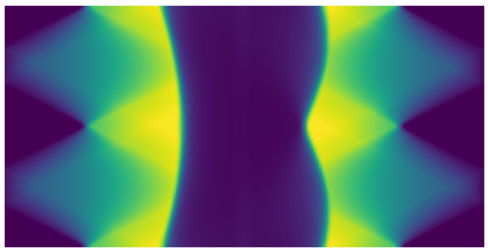


Figure 19: First estimated of MLEM $LA\lambda_m + r$

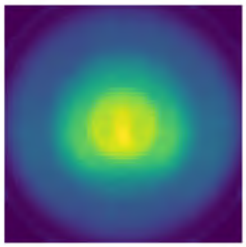


Figure 20: Back Projection of Relative distance $A^T L^T \frac{M}{LA\lambda+r}$

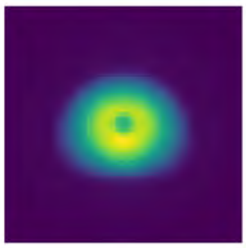


Figure 21: First estimated of MLEM using equation 12

Figure 22 demonstrates the initialization and subsequent reconstructed activity maps after 25, 50, and the final iteration. This sequential depiction highlights the iterative improvement in image quality and convergence towards an accurate representation of the activity distribution.

For the scatter correction, we need to compute the scaling factor, which represents the multiplication term

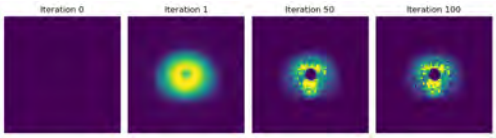



Figure 22: The reconstructed activity map of MLEM reconstuction without scatter correction

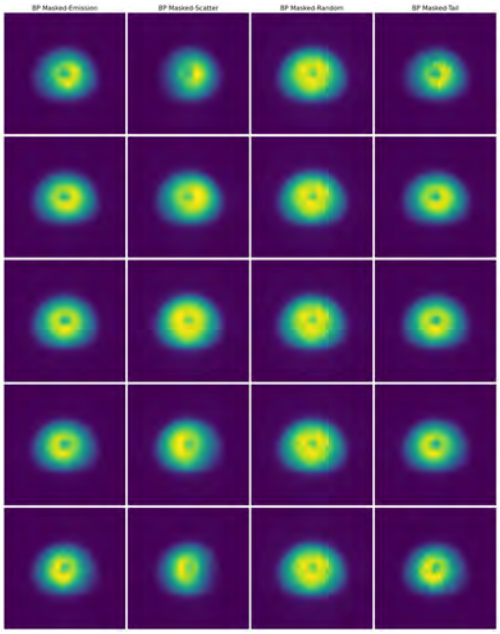


Figure 23: The attribution to compute the scaling factor. The row represent each TOF Bin

applied to each bin of the scatter sinogram in the tail sinogram. The tail sinogram is computed as explained in the Methodology section. This scaling factor is critical for accurately estimating and correcting scatter contributions in the reconstructed images.

To illustrate the scatter profile after multiplication with the scaling factor, the scatter profile compared with the emission sinogram is shown in Figure 26. Note that this scatter sinogram is simulated from the activity and attenuation map and used in the emission data before adding Poisson noise, as represented by the fractured graph in the mentioned figure. This comparison highlights the impact of scatter correction on the overall image quality and accuracy.

Subsequently, the MLEM algorithm incorporating the scatter term is computed, and the results are presented in Figure 24. This figure showcases the improved reconstruction quality achieved by integrating scatter correction into the MLEM process.

However, this approach is primarily used to establish a baseline performance, as we do not have the initial activity image. This necessitates performing reconstruction for several iterations to obtain the reconstructed activity map, which is then used to generate the scatter sinogram. The results of this algorithm, which gener-

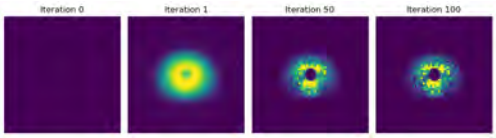



Figure 24: The reconstructed activity map of MLEM reconstuction with static scatter correction

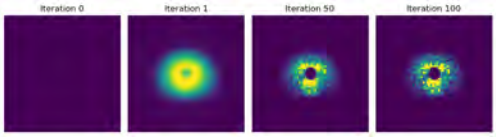



Figure 25: The reconstructed activity map of MLEM reconstuction with scatter correction using iterative scatter simulation

ates a new scatter sinogram every 10 epochs of MLEM, are shown in Figure 25. Nonetheless, the scatter sinogram profile from the final update is depicted in Figure 26, providing a comprehensive view of the iterative improvement in scatter correction.

At the end of the evaluation in synthetic data, Log loss, as desired for the MLEM algorithm, PSNR and SSIM for image comparison are used as evaluation scores to observe the performance of the reconstruction task. These scores are evaluated in the image space since we have ground truth.

The log loss comparison is shown in Figure 27, evaluating the performance in terms of the logarithmic loss function. The MLEM method (black crosses) exhibits a gradual decrease in log loss, stabilizing around 2430. The baseline scatter correction (red circles) demonstrates a sharper decline, stabilizing at a lower log loss of approximately 2415. The iterative scatter estimation (yellow squares) shows the most rapid decrease initially, but with higher variability, ultimately stabilizing close to the baseline scatter correction at around 2415. This suggests that both the baseline and iterative scatter correction methods offer significant improvements in reducing log loss compared to the MLEM method, with

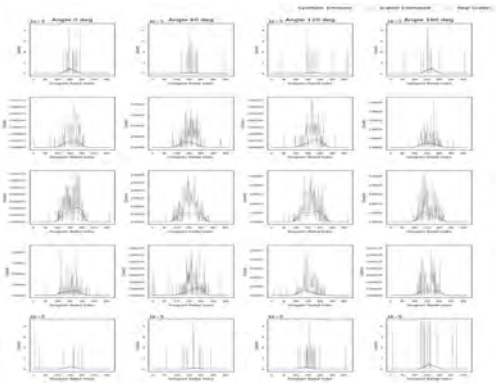


Figure 26: Scaling profile of scaled computed scatter sinogram compare to emission data and scatter sinogram used to generate emission data

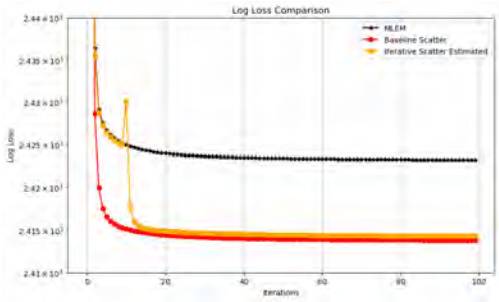



Figure 27: Log Loss Comparison across 100 iterations for MLEM, baseline scatter correction, and iterative scatter estimation methods.

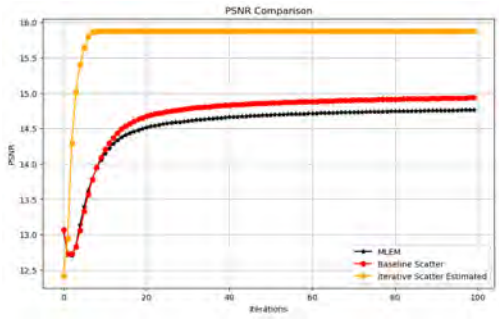



Figure 28: PSNR Comparison across 100 iterations for MLEM, baseline scatter correction, and iterative scatter estimation methods.

the iterative method showing more variability but comparable final performance.

The Peak Signal-to-Noise Ratio (PSNR) comparison is depicted in Figure 28. This graph illustrates the PSNR values across 100 iterations for three methods: Maximum Likelihood Expectation Maximization (MLEM), baseline scatter correction, and iterative scatter estimation. The MLEM method, represented by black crosses, shows a steady increase in PSNR, stabilizing around 14.5 dB. The baseline scatter correction, depicted with red circles, demonstrates a more rapid increase in PSNR, reaching approximately 15 dB, indicating a noticeable improvement over the MLEM. The iterative scatter estimation method, shown in yellow squares, achieves the highest PSNR values, quickly rising to about 16 dB and maintaining this level throughout the iterations. This suggests that the iterative scatter estimation method significantly enhances image quality by effectively reducing noise and improving signal accuracy.

Figure 29 presents the Structural Similarity Index (SSIM) comparison, which measures the similarity between two images. The SSIM values are plotted for the same three methods over 100 iterations. The MLEM method (black crosses) and baseline scatter correction (red circles) show similar SSIM values, with both methods stabilizing around 0.6. The iterative scatter estimation (yellow squares), however, starts lower but gradually increases, plateauing just below 0.6. Although the iterative method improves over time, it does not surpass

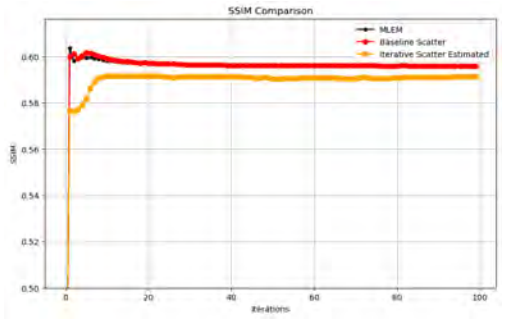



Figure 29: SSIM Comparison across 100 iterations for MLEM, baseline scatter correction, and iterative scatter estimation methods.

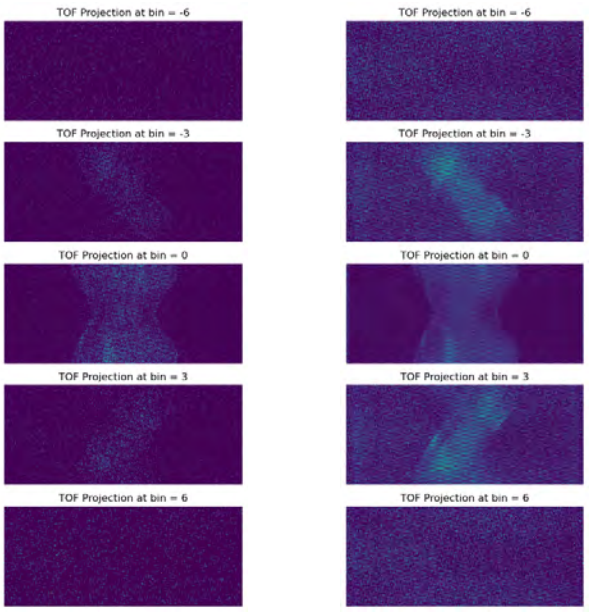
the baseline scatter correction in terms of SSIM, possibly due to the noise and limited scanner size used in this study. This indicates that while the iterative method excels in PSNR, its SSIM performance is less pronounced, suggesting a trade-off between different image quality metrics.

In summary with synthetic data experiment, the iterative scatter estimation method demonstrates superior performance in terms of PSNR and log loss, indicating better overall image quality and reduced noise. However, its SSIM improvement is less pronounced, highlighting the complexity of balancing different image quality metrics in PET imaging. Further optimization and testing with larger datasets and real clinical scenarios will be necessary to fully validate these findings.

To observe the performance in real clinical data, the data acquired using a Siemens mCT scanner is used. The scanner has 13 TOF bins with 55 rings and 672 detectors. We compare the results after using MLAA with and without scatter correction. The table presents a comparative analysis between the MLAA and MLAA with scatter correction (MLAA_S) methods, focusing on computational time, SNR in Image space while SSIM and PSNR for Sinogram space.

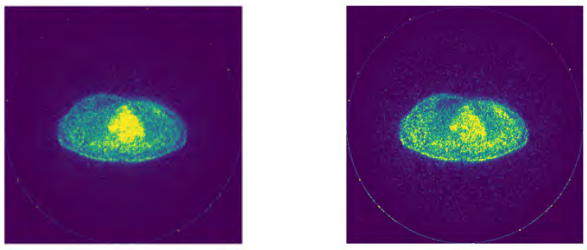
For the experiment with 20 iteration, the MLAA method demonstrated a computational time of 226 minutes, yielding an SNR of 0.2455, a PSNR of 28.7802, and an SSIM of 0.3697. In contrast, the MLAA_S method required a longer computational time of 268 minutes and produced a slightly lower SNR of 0.1902, a marginally reduced PSNR of 28.7200, and a slightly decreased SSIM of 0.3654. These results suggest that while the scatter correction in MLAA_S increases the computational burden, it does not significantly enhance image quality metrics compared to the standard MLAA method.

The comparison of reconstructed activity maps is shown in Figure 31. Figure 32 show the result from scatter simulation using the activity and attenuation maps from the MLAA reconstruction process in singram space. In the other hand, the effect of scattering in image space are presented in Figures 33, 34, and 35,



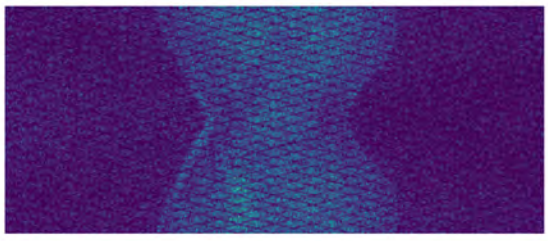

(a) Emission Sinogram (b) Estimated Emission Sinogram

Figure 30: Comparison of Measured Prompts and Estimated Emission Prompts using MLAA with Scatter Correction after 20 iterations

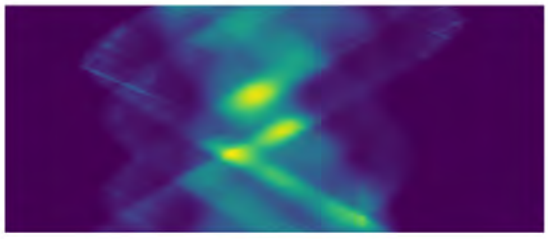


(a) MLAA Reconstuction (b) MLAA_S Reconstruction

Figure 31: Comparison of estimated activity distribution using MLAA with and without scatter correction after 20 iterations



(a) Non-TOF Emission prompts



(b) Non-TOF Scattering prompts

Figure 32: Scattering effect in the coronal views

| Method | SNR | PSNR (dB) | SSIM | Time (m.) |
|---|---|---|---|---|
| MLAA | 0.24 | 28.78 | 0.37 | 226 |
| MLAA_S | 0.19 | 28.72 | 0.37 | 268 |

Table 2: Comparison of MLAA and MLAA with scatter correction (MLAA_S).

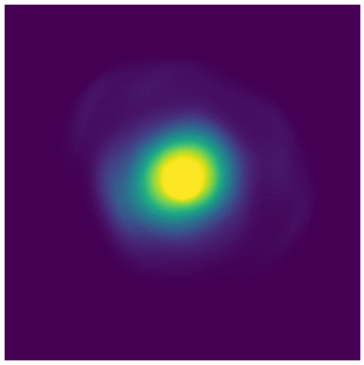
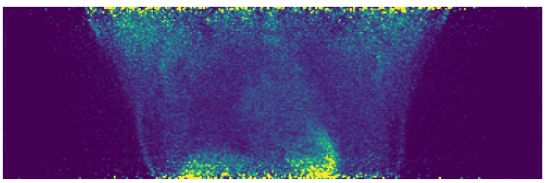


Figure 33: Scattering effect to the prompts.

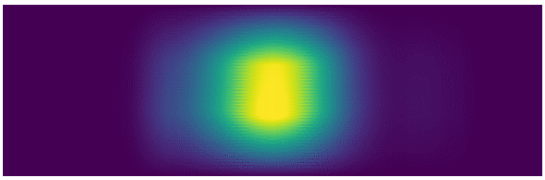
respectively, each depicting different views.

## 5. Discussion

The implementation of the Time-of-Flight (TOF) single scatter simulation (SSS) was successfully achieved, demonstrating significant improvements in computational efficiency. This efficiency gain substantially reduced computational time, making the simulation more practical for large-scale applications. However, ongoing tests with real clinical data and synthetic data generated through Monte Carlo simulations are in progress in collaboration with the openSSS developer community. These tests aim to further validate the implementation under diverse conditions and with more complex datasets.

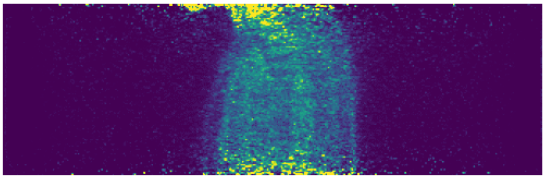


(a) MLAA with scatter correction

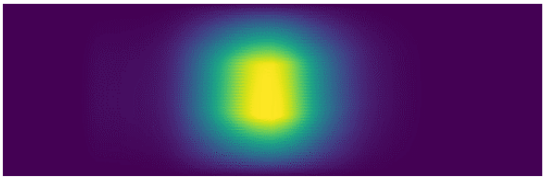


(b) Scattering effect

Figure 34: Scattering effect in the coronal views



(a) MLAA with scatter correction



(b) Scattering effect

Figure 35: Scattering effect in the sagital views

For our internal validation, the implemented SSS was used to generate data and run scatter correction using the generated scatter sinogram. The complete pipeline was evaluated in both baseline correction scenarios, where the scatter simulation was used to generate the emission sinogram itself, and in scenarios where scatter sinograms were generated from reconstructed sinograms. This dual approach provided a comprehensive assessment of the scatter correction's performance.

The results of the scatter correction were promising, showing good performance in terms of Peak Signal-to-Noise Ratio (PSNR) and log loss metrics. However, the Structural Similarity Index (SSIM) did not show significant improvement. This could be attributed to the strong presence of noise and the relatively small size of the scanner used in this study. The small geometry, while advantageous for saving computational time, might limit the SSIM improvement. Despite this, the scatter-corrected results still showed overall better image quality.

For the clinical data acquired using the Siemens mCT scanner, the comparison between the MLAA and MLAA with scatter correction. As expected that the computational time for the MLAA with Scatter Correction method was longer, requiring 40 minutes longer. The SNR for the MLAA method was higher at 0.2455 compared to 0.1902 for the MLAA with scatter correction method, while PSNR and SSIM is comparable. These results indicate that while the scatter correction slightly increased the computational burden, it did not significantly improve the image quality metrics (SNR, PSNR, SSIM). The negligible improvement in SSIM suggests that the noise level and small geometry of the scanner continue to play a significant role in limiting image quality enhancement. However, the overall image quality with scatter correction still showed better performance, reinforcing the benefits of implementing scatter correction in clinical settings despite the additional computational cost.

Nevertheless, the practical application of this scatter simulation technique poses challenges. The iterative reconstruction process, which updates the reconstructed image or activity map, leads to high computational demands, particularly for real-world scenarios involving large-scale scanners like the Siemens mCT scanner with 55 rings and 672 detectors, or the Siemens Vision 600 with 80 rings and 760 detectors. Even with the faster implementation developed in this study, the computational load remains substantial, suggesting that this approach might not be ideal for practical, routine use.

To address these challenges, further optimization is necessary. This includes finding the optimal size for activity and attenuation images, selecting appropriate samples of detectors and rings, and possibly reducing the number of scatter points in the simulation. These optimizations could significantly enhance the feasibility of the simulation for practical applications.

In parallel, exploring advanced approaches such as generating scatter sinograms based on activity images and attenuation maps using deep learning techniques could offer a promising direction for future work. Deep learning models have the potential to learn complex patterns and relationships in the data, potentially providing a more efficient and accurate method for scatter correction.

## 6. Conclusions

This study successfully developed and implemented a Single Scatter Simulation (SSS) algorithm using Python, optimized with Numba and parallel processing, achieving significant reductions in computational time for three-dimensional positron emission tomography (PET) imaging. Validation against MATLAB benchmarks confirmed the enhanced efficiency of the Python implementation, making it feasible for large-scale applications.

The integration of SSS results into iterative reconstruction techniques, specifically the Maximum Likelihood Reconstruction of Attenuation and Activity (MLAA), and the incorporation of Time-of-Flight (TOF) information, showed promising improvements in image quality. Evaluations with synthetic and real clinical data demonstrated good performance in terms of Peak Signal-to-Noise Ratio (PSNR) and log loss metrics, although Structural Similarity Index (SSIM) improvements were limited due to noise and the small scanner size.

Despite these advancements, the high computational demands of the iterative reconstruction process for large-scale scanners pose practical challenges. Future research should focus on optimizing image sizes, detector and ring samples, and scatter point reduction. Additionally, exploring deep learning techniques for generating scatter sinograms offers a promising direction for enhancing scatter correction methods, potentially revolutionizing PET image reconstruction and improving diagnostic outcomes.

## Acknowledgments

First and foremost, I extend my deepest gratitude to my supervisor, Massimiliano Colarieti Tosti, for welcoming me into the lab and providing invaluable guidance and support. This opportunity has allowed me to learn and grow significantly in my academic journey.

I am profoundly thankful to the MAIA program and the Erasmus Mundus scholarship for their financial and academic support throughout these two years of my master's degree. Their assistance has been crucial in enabling me to pursue and achieve my educational goals.

I also wish to express my heartfelt appreciation to my family, friends, and colleagues for their unwavering support and encouragement. Your belief in me has been a source of strength and motivation.

Lastly, I would like to acknowledge myself for the hard work and dedication that I have invested in my studies. This accomplishment would not have been possible without perseverance and self-discipline.

## References

Barney, J., Rogers, J., Harrop, R., Hoverath, H., 1991. Object shape dependent scatter simulations for pet. IEEE Transactions on Nuclear Science 38, 719–725. doi:10.1109/23.289380.

Benoit, D., Ladefoged, C.N., Rezaei, A., Keller, S.H., Andersen, F.L., Højgaard, L., Hansen, A.E., Holm, S., Nuyts, J., 2016. Optimized mlaa for quantitative non-tof pet/mr of the brain. Physics in Medicine & Biology 61, 8854.

Beyer, T., Townsend, D.W., Brun, T., Kinahan, P.E., Charron, M., Roddy, R., Jerin, J., Young, J., Byars, L., Nutt, R., 2000. A combined pet/ct scanner for clinical oncology. Journal of nuclear medicine 41, 1369–1379.

Budinger, T.F., 1983. Time-of-flight positron emission tomography: status relative to conventional pet.

Buvat, I., Benali, H., Todd-Pokropek, A., Di Paola, R., 1994. Scatter correction in scintigraphy: the state of the art. European Journal of Nuclear Medicine 21, 675–694.

Cherry, S.R., Sorenson, J.A., Phelps, M.E., 2013. Physics in nuclear medicine. Soc Nuclear Med.

Deans, S.R., 2007. The Radon transform and some of its applications. Courier Corporation.

Goggin, A., Ollinger, J., 1994. A model for multiple scatters in fully 3d pet, in: Proceedings of 1994 IEEE Nuclear Science Symposium - NSS'94, pp. 1609–1613 vol.4. doi:10.1109/NSSMIC.1994.474755.

Hermanns, C.F., 2013. X-ray absorption studies of metalloporphyrin molecules on surfaces: Electronic interactions, magnetic coupling, and chemical switches. Ph.D. thesis. Freie Universität Berlin.

Hill, C., 2019. The klein–nishina formula. URL: `https://scipython.com/blog/the-kleinnishina-formula`. accessed: 2024-06-06.

Holdsworth, C., Badawi, R., Santos, P., Van den Abbeele, A., Hoffman, E., El Fakhri, G., 2003. Evaluation of a monte carlo scatter correction in clinical 3d pet, in: 2003 IEEE Nuclear Science Symposium. Conference Record (IEEE Cat. No.03CH37515), pp. 2540–2544 Vol.4. doi:10.1109/NSSMIC.2003.1352408.

Hutton, B.F., Nuyts, J., Zaidi, H., 2006. Iterative Reconstruction Methods. Springer US, Boston, MA. pp. 107–140. doi:10.1007/0-387-25444-7_4.

Jiang, W., Chalich, Y., Deen, M.J., 2019. Sensors for positron emission tomography applications. Sensors 19, 5019.

Kak, A.C., Slaney, M., 2001. Principles of computerized tomographic imaging. SIAM.

Klein, O., Nishina, Y., 1928. The scattering of light by free electrons according to dirac's new relativistic dynamics. Nature 122, 398–399.

Lange, K., Carson, R., 1984. Em reconstruction algorithms for emission and transmission tomography. Journal of computer assisted tomography 8, 306–316.

Levin, C., Dahlbom, M., Hoffman, E., 1995. A monte carlo correction for the effect of compton scattering in 3-d pet brain imaging. IEEE Transactions on Nuclear Science 42, 1181–1185. doi:10.1109/23.467880.

Lewitt, R., Matej, S., 2003. Overview of methods for image reconstruction from projections in emission computed tomography. Proceedings of the IEEE 91, 1588–1611. doi:10.1109/JPROC.2003.817882.

Manglos, S.H., Gagne, G.M., Krol, A., Thomas, F.D., Narayanaswamy, R., 1995. Transmission maximum-likelihood reconstruction with ordered subsets for cone beam ct. Physics in Medicine & Biology 40, 1225. URL: https://dx.doi.org/10.1088/0031-9155/40/7/006, doi:10.1088/0031-9155/40/7/006.

Natterer, F., Wübbeling, F., 2001. Mathematical methods in image reconstruction. SIAM.

Ollinger, J., Johns, G., 1993. Model-based scatter correction for fully 3d pet, in: 1993 IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference, pp. 1264–1268. doi:10.1109/NSSMIC.1993.701845.

Panin, V.Y., 2012. Scatter estimation scaling with all count use by employing discrete data consistency conditions, in: 2012 IEEE Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC), pp. 2998–3004. doi:10.1109/NSSMIC.2012.6551685.

Santo, R., Salomon, A., Jong, H., Stute, S., Merlin, T., Beijst, C., 2023. opensss: an open-source implementation of scatter estimation for 3d tof-pet, pp. 1–1. doi:10.1109/NSSMICRTSD49126.2023.10338342.

Seibert, J.A., Boone, J.M., 2005. X-ray imaging physics for nuclear medicine technologists. part 2: X-ray interactions and image formation. Journal of nuclear medicine technology 33, 3–18.

Sheikh, H., Bovik, A., Veciana, G., 2006. An information fidelity criterion for image quality assessment using natural scene statistics. Image Processing, IEEE Transactions on 14, 2117 – 2128. doi:10.1109/TIP.2005.859389.

Shepp, L.A., Vardi, Y., 1982. Maximum likelihood reconstruction for emission tomography. IEEE Transactions on Medical Imaging 1, 113–122. doi:10.1109/TMI.1982.4307558.

Thielemans, K., Manjeshwar, R., Tsoumpas, C., Jansen, F., 2007. A new algorithm for scaling of pet scatter estimates using all coincidence events, in: 2007 IEEE Nuclear Science Symposium Conference Record, pp. 3586–3590. doi:10.1109/NSSMIC.2007.4436900.

Vandenberghe, S., Moskal, P., Karp, J.S., 2020. State of the art in total body pet. EJNMMI physics 7, 1–33.

Verel, I., Visser, G.W., Van Dongen, G.A., 2005. The promise of immuno-pet in radioimmunotherapy. Journal of Nuclear Medicine 46, 164S–171S.

Wang, Z., Bovik, A., 2002. A universal image quality index. IEEE Signal Processing Letters 9, 81–84. doi:10.1109/97.995823.

Watson, C., 1999. New, faster, image-based scatter correction for 3d pet, in: 1999 IEEE Nuclear Science Symposium. Conference Record. 1999 Nuclear Science Symposium and Medical Imaging Conference (Cat. No.99CH37019), pp. 1637–1641 vol.3. doi:10.1109/NSSMIC.1999.842888.

Watson, C., 2005. Extension of single scatter simulation to scatter correction of time-of-flight pet, in: IEEE Nuclear Science Symposium Conference Record, 2005, pp. 2492–2496. doi:10.1109/NSSMIC.2005.1596846.

Watson, C., Casey, M., Michel, C., Bendriem, B., 2004. Advances in scatter correction for 3d pet/ct, in: IEEE Symposium Conference Record Nuclear Science 2004., pp. 3008–3012. doi:10.1109/NSSMIC.2004.1466317.

Watson, C.C., Newport, D., Casey, M.E., 1996. A Single Scatter Simulation Technique for Scatter Correction in 3D PET. Springer Netherlands, Dordrecht. pp. 255–268. doi:10.1007/978-94-015-8749-5_18.

Zaidi, H., Koral, K.F., 2004. Scatter modelling and compensation in emission tomography. European Journal of Nuclear Medicine and Molecular Imaging 31, 761–782. URL: https://api.semanticscholar.org/CorpusID:2838946.

Zuo, C., Li, J., Sun, J., Fan, Y., Zhang, J., Lu, L., Zhang, R., Wang, B., Huang, L., Chen, Q., 2020. Transport of intensity equation: a tutorial. Optics and Lasers in Engineering 135, 106187.

## Appendix

Table 3: ToyScanner Geometry

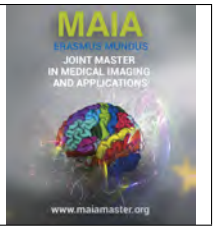
| Name | Value |
|---|---|
| DetectorSize | [5,5] |
| EnergyResolution | 0.1500 |
| NrCrystalsAxial | 4 |
| NrCrystalsTrans | 4 |
| NrModulesAxial | 2 |
| NrModulesTrans | 2 |
| NrSectorsAxial | 1 |
| NrSectorsTrans | 40 |
| TOFResolution | 400 |

Table 4: Siemens_mcT Geometry

| Name | Value |
|---|---|
| DetectorSize | [4,4] |
| EnergyResolution | 0.1150 |
| NrCrystalsAxial | 14 |
| NrCrystalsTrans | 14 |
| NrModulesAxial | 4 |
| NrModulesTrans | 1 |
| NrSectorsAxial | 1 |
| NrSectorsTrans | 48 |
| TOFResolution | 555 |

Medical Imaging and Applications

Master Thesis, June 2024




# Deep learning-based survival prediction for pancreatic cancer using histopathology images

Jaqueline A. Leal Castillo, Noémie Moreau, Katarzyna Bozek

*Center for Molecular Medicine Cologne, Germany*

**Abstract**

Pancreatic ductal adenocarcinoma (PDAC) is the most lethal form of pancreatic cancer. Accurate survival prediction could guide treatment strategies and facilitate patient stratification for critical trials. A significant challenge is the complexity of gigapixel Whole Slide Images (WSIs), which contain vast amounts of detailed information, making it difficult for current deep learning methods to accurately analyze these structures. This study explores the application of GraphLSurv for the prediction of survival in the context of PDAC. GraphLSurv is a scalable Graph Convolution Network (GCN) designed for survival prediction from Whole Slide Images. It dynamically generates adaptive and sparse graph structures to highlight and analyze important areas within WSIs and Tissue Microarrays (TMAs). Two different models were used as feature extractors to generate the input of the network: ResNet-50 pretrained on ImageNet, and a vision transformer backbone (ViT-L/16 via DINOv2) pretrained on histopathology-specific datasets. Our method was tested on a PDAC dataset from the University of Cologne, including 850 WSIs, 650 TMAs, and survival information from 321 patients. The results show that GraphLSurv is able to predict patients risk in the context of PDAC. With a concordance index of 0.5984, we showed the model using features extracted from WSIs using the vision transformer backbone (ViT-L/16 via DINOv2) pretrained on histopathology-specific datasets was more performant. This underscores the importance of domain-specific feature extraction in developing robust prognostic models. In the end, our best model was able to separate the patient in two groups: high-risk and low-risk with p-value equal to 0.0410 for a log-rank test. The implications of this study presents that by combining advanced deep learning techniques with specialized feature extraction, GraphLSurv represents a major step forward in creating more accurate and reliable survival prediction models for PDAC.

*Keywords:* Survival Prediction, Pancreatic cancer, Graph Convolutional Network, Whole Slide Images, Tissue Microarrays

## 1. Introduction

Pancreatic cancer, particularly pancreatic ductal adenocarcinoma (PDAC), is the seventh leading cause of cancer death worldwide. Characterized by its high lethality, PDAC has a five-year survival rate of around 4% for the most common form, malignancy of the exocrine pancreas (Hidalgo et al., 2015). At the time of diagnosis, only 15-20% of patients are eligible for surgical intervention, which itself only slightly improves the five-year survival rate to 20% (Mizrahi et al., 2020). The insidious nature of pancreatic cancer, marked by few and nonspecific symptoms, complicates early detection, leaving the majority of patients with limited treatment options. Survival analysis is critical for understanding and improving outcomes in pancreatic cancer. It involves methods to assess the probability of survival from the time of diagnosis to a specific future time, the hazard or the risk of the event, and the survival rate of patients not experiencing the event after a certain period (Clark et al., 2003). Traditional survival analysis approaches, such as Cox proportional hazards models, offer insights into associations between patient variables and survival outcomes. However, they often rely on univariate and multivariable regression analyses that may not capture the complex interactions within high-dimensional data, such as histopathological im-

ages (Bradburn et al., 2003).

Histopathological Whole-Slide Images (WSIs) are a type of medical pathology image typically used by pathologists to diagnose complex tumor diseases, including tumor invasion, mitosis, anaplasia, and necrosis. These images enable clinical doctors to make critical decisions on disease treatment. The advent of Whole Slide Imaging (WSI) has revolutionized the field of digital pathology, offering numerous advantages over traditional microscopy. Over the past decade, advancements in image digitizing technology have led to the development of slide scanners capable of producing WSIs, which can be examined similarly to conventional microscopy but with added benefits. WSIs allow for high-resolution capture and detailed examination of tissue samples, crucial for accurate diagnoses and research. They provide consistent quality over time and facilitate the use of image processing techniques, enhancing the diagnostic process. Moreover, WSIs can be easily shared and accessed remotely, making them invaluable for telepathology, education, and collaborative research (Al-Janabi et al., 2012).

Tissue Microarrays (TMAs) are crucial in digital pathology, enabling the simultaneous analysis of multiple samples, reducing costs, and enhancing tumor profiling standardization. TMAs are constructed by taking small cylindrical cores of tissue samples from different tumor specimens or patients and arranging them in an array pattern on a single recipient paraffin block. The figure 1 illustrates the array of tissues and a close-up on the individual cores. This approach minimizes staining variations and improves diagnostic reliability. Digital TMA slides optimize documentation, storage, and result retrieval. Implementing TMAs in diagnostics offers a reliable and economical method for tumor classification, essential for personalized treatment strategies (Rossing et al., 2012).

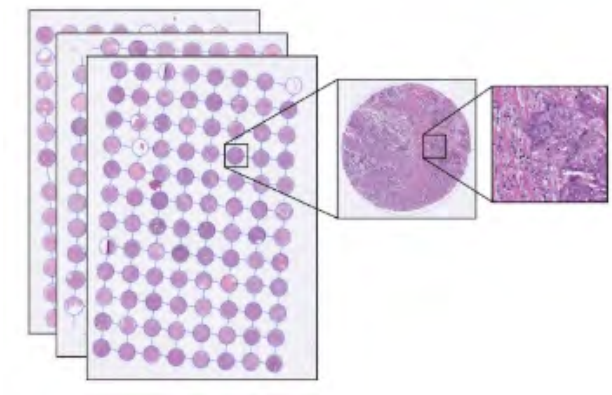


Figure 1: Magnified view of a tissue patch extracted from one core of a TMA (Sandarenu et al., 2022).

Despite these advancements, manual interpretation of histopathological images remains subjective, suffering from large inter- and intra-observer variability. Even patients with the same histopathological features can have distinct survival outcomes due to tumor heterogeneity (Ren et al., 2022). In recent years, deep learning has significantly advanced computational histopathology. These advancements enable the extraction of clinically useful biomarkers directly from Whole Slide Images (WSIs), enhancing cancer prognosis by integrating extensive histological data. This integration leads to improved survival prediction and treatment outcomes while also augmenting the expertise of pathologists. Notably, deep learning techniques automate tasks such as tumor detection, grading, and subtyping with high accuracy, often surpassing pathologist-level performance (Cooper et al., 2023). Deep learning models can predict genetic mutations and survival outcomes from histopathology images, offering valuable prognostic insights. Techniques like multiple-instance learning (MIL), transfer learning, and weakly-supervised learning enable effective analysis of large datasets, enhancing model performance even with limited labeled data. Weakly-supervised learning trains models on complex tasks using image-level labels or noisy annotations, making tumor detection possible without detailed annotations (Li et al., 2023). Self-supervised learning methods, help models learn rich representations from unlabeled data. Neural attention mechanisms improve interpretability and performance by focusing on relevant regions in histopathology images. These advancements have facilitated personalized treatment plans and reduced the burden on pathologists, helping for more precise and efficient cancer diagnostics and prognostication.

Graph Convolutional Networks (GCNs) provide a promising method by representing histopathological data as graphs, where nodes stand for tissue patches and edges show the relationships between these patches. This approach offers a deeper understanding of the tumor microenvironment, which is crucial for cancer progression and patient outcomes. Unlike traditional CNNs, GCNs can dynamically adjust to the spatial relationships within the data, providing a more comprehensive and adaptive analysis (Wang et al., 2022).

### *1.1. Key Contributions*

Our contributions in this research are summarized as follows:

- Exploration of GraphLSurv: This project applies GraphLSurv, a weakly supervised survival prediction framework, to a specific pancreatic ductal adenocarcinoma (PDAC) dataset, allowing for direct predictions of patient-level outcomes without relying on local labels.
- Feature extraction comparison: The study employs a novel feature extractor, ViT/16, pretrained on histopathology images, and compares its performance with the traditional ResNet-50 model. This

comparison highlights the importance of tailored pretraining for histopathological tasks.

- Integration of Tissue Microarrays (TMAs): In contrast to the predominant focus on WSI, this thesis incorporates TMAs, providing a fresh perspective on histopathological analysis. TMAs offer a standardized and cost-effective means to study multiple tissue samples, complementing insights gained from WSIs.

## 2. State of the art

### *2.1. Survival Analysis*

In cancer research, the focus is often on the time until a specific event, such as relapse or death, occurs, known as survival time. Since not all individuals experience the event during the study period and survival data are typically skewed with many early events, specialized methods known as survival analysis are used. Survival analysis assesses the probability of survival from diagnosis to a future time, the risk of the event, and the survival rate (Clark et al., 2003).

The traditional Cox model, a widely used method in medical research for analyzing survival data, investigates how factors like age or treatment affect the risk of events like death over time (Bradburn et al., 2003). It assumes these factors consistently influence event rates, reflected in hazard ratios. Validating the proportionality assumption is crucial for the model's reliability in predicting survival outcomes. Essentially, the Cox model employs multiple linear regression of the hazard's logarithm on variables $x_i$, with the baseline hazard varying over time. Covariates then multiply the hazard at any time point, reflecting the key proportional hazards model assumption: event hazards in different groups are constant multiples of each other. Another popular method for estimating survival probability is Kaplan-Meier (KM) (Bradburn et al., 2003), which accounts for both censored and uncensored survival times. For $k$ patients that have events at distinct times $t_1 < t_2 < t_3 < \ldots < t_k$. The KM method calculates the probability of surviving each time interval and multiplies these probabilities to get the overall survival probability. Each patient contributes information until they have an event or are censored. If no patients were censored, the KM estimate would be the number of event-free individuals at time $t$ divided by the total number of study participants. Commonly used evaluation metrics for survival models include the Concordance Index (C-Index). The C-Index measures the model's ability to correctly rank survival times based on the predicted risk scores. It is defined as the probability that, for a randomly selected pair of subjects, the subject with the higher predicted risk score experiences the event before the subject with the lower predicted risk score. The C-Index ranges from 0.5 for random predictions to 1, where 1 indicates perfect discrimination. A higher C-Index value indicates better discriminative ability of the prognostic model.

### *2.2. Survival Analysis with Deep Learning*

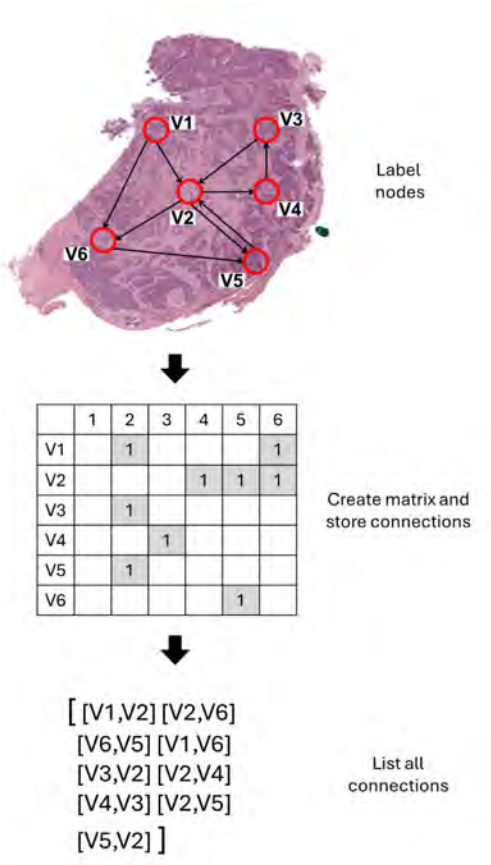
Traditional survival analysis methods, like those mentioned above, have been used to analyze and model survival data. However, with the rise of machine learning and deep learning, new approaches have emerged that include these techniques to improve survival prediction. The survival prediction problem can be formulated as follows:

$$\hat{y}_i(B_i) = F(B_i; \theta) \tag{1}$$

where $\hat{y}_i$ is the predicted survival time or event time for the $i$-th individual, $B_i$ represents the set of covariates or features, such as age, gender, treatment, etc. $F$ is a function that maps the covariates $B_i$ to the predicted survival time $\hat{y}_i$, parameterized by $\theta$, that denotes the parameters of the survival prediction model, learned from training data. The goal is to estimate the function $F$ and its parameters $\theta$ using historical data, which includes both censored and uncensored observations. Censored observations refer to instances where the event of interest has not been observed during the study period. The Cox proportional hazards model is used as the survival loss function, optimizing the network to predict survival risk scores directly from the graph representations (Baek et al., 2021). The negative log-likelihood function or the Cox model is given by:

$$\ell_{\text{cox}} = \sum_{i \in \{i:\sigma_i=1\}} \left( \hat{y}_i - \log \sum_{j \in \{j:t_j \geq t_i\}} e^{\hat{y}_j} \right)$$

where $\hat{y}_i$ is the risk score for patient $i$, and $\sigma_i$ is the set of patients still at risk at time $t_i$. DeepSurv enhances the Cox model using a neural network to capture nonlinear covariate effects, outperforming traditional methods (Katzman et al., 2018). DeepConvSurv, a deep convolutional survival model for predicting survival from histopathological images, operates at the patch level with a unique architecture utilizing Cox model loss. It aggregates patch-level risks to derive patient-level predictions (Zhu et al., 2016). WSISA (Whole Slide Imaging Survival Analysis) stands as a comprehensive survival prediction framework for WSIs, incorporating patch sampling, clustering, and cluster selection stages (Zhu et al., 2017). It showcases robustness in handling variability in WSI sizes and patterns. DeepAttnMISL introduces an attention-based aggregation approach for survival prediction using WSIs, emphasizing the importance of relevant regions via an attention mechanism (Yao et al., 2020).




| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| V1 | | 1 | | | | 1 |
| V2 | | | | 1 | 1 | 1 |
| V3 | | 1 | | | | |
| V4 | | | 1 | | | |
| V5 | | 1 | | | | |
| V6 | | | | | 1 | |

Figure 2: Visual depiction illustrating the process of constructing a Graph structure from a Whole Slide Image (WSI). It begins with the establishment of labeled nodes, with each patch depicted as a node, interconnected by edges representing their connections.

### 2.3. Graph Convolution Networks (GCN) for Survival Analysis

Graph Convolutional Networks (GCNs) are neural networks designed for graph-structured data. Unlike traditional neural networks that handle Euclidean data such as images or text, GCNs manage non-Euclidean structures represented as graphs. They extend convolutional operations from regular Convolutional Neural Networks (CNNs) to graph data, aggregating information from neighboring nodes to update node features. In GCNs, each node in the graph has its own feature vector, which is updated by aggregating the feature vectors of neighboring nodes. This aggregation process combines neighboring node features and applies transformations to generate updated node feature vectors. Mathematically, a single graph convolutional layer operation can be expressed as a formula involving node feature matrices and weight matrices. The graph convolutional network equation is defined as:

$$H^{(l+1)} = \sigma\left(D^{-\frac{1}{2}}AD^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)$$

where $H^{(l)}$ is the node feature matrix at layer $l$, $A$ is the adjacency matrix, $D$ is the degree matrix, $W^{(l)}$ is the layer-specific trainable weight matrix, and $\sigma$ is the activation function. For tasks requiring an overall understanding of the graph, such as graph classification, GCNs aggregate node features into a single representation of the entire graph using pooling operations like mean or max pooling. GCNs offer several advantages for survival analysis by capturing complex relationships between variables. Survival data often involves intricate connections between variables, such as patient demographics, clinical features, and genetic data. GCNs model these relationships by representing the data as a graph where nodes represent patients and edges indicate relationships, such as similarity in clinical features. In the case of WSI as represented in 2 the nodes will be represented as the patches and the edges as the relationships between them. By aggregating information from neighboring nodes, GCNs can effectively reduce dimensionality and capture relevant patterns, also can improve the predictive performance of survival models by capturing non-linear interactions and dependencies that traditional methods might not be able to capture.

### 2.4. Feature Extraction from WSIs

Developing machine learning-based diagnostic models for histopathology images has gained significant attention. Deep learning models trained on natural images often underperform on histopathology images due to domain shift, emphasizing the need for specialized feature extractors. While pre-trained models like ResNet, VGG, or Inception are widely used, they may not capture the unique features required for specific fields such as medical imaging. Therefore, there is increasing interest in creating specialized feature extractors by pre-training models on domain-specific datasets. Despite these challenges, several strategies are being explored to address data scarcity. These include data augmentation, transfer learning, self-supervised or weakly-supervised learning, and multi-center collaborations that merge resources to create larger, more diverse datasets (Xu et al., 2022). However, models trained on limited or public datasets may not generalize well to unseen data from different demographics or imaging protocols, further limiting their practical utility (Li et al., 2020). The usage of CLAM (ResNet-50) (Lu et al., 2021) pretrained on ImageNet is popular in the survival prediction projects, such as: PatchGCN, designed for histopathology images, uses graph representations of image patches to predict cancer survival outcomes, improving prediction accuracy by leveraging spatial relationships between patches (Chen et al., 2021). GraphLSurv, a scalable graph convolution network, predicts survival from gigapixel Whole-Slide Images (WSIs). It overcomes the limitations of traditional methods by generating adaptive, sparse structures for patches, dynamically capturing latent correlations, and reducing computational complexity with an anchor-based technique (Liu et al., 2023a). Furthermore, ex-

ploring models that comprises over 100 million histology images from diverse sources, such as UNI (Chen et al., 2024), presents an intriguing opportunity for further investigation.

## 3. Material and methods

### *3.1. Dataset*

This study employs a unique dataset of pancreatic ductal adenocarcinoma (PDAC) with histopathology images provided by the University of Cologne as part of the PANCALYZE trial protocol (Popp et al., 2017). The dataset includes Whole Slide Images (WSIs) and Tissue Microarrays (TMAs) at 40x magnification from various regions of the pancreas (head, body, tail), along with clinical data such as recurrence dates, survival times and censorship status. The dataset comprises 850 WSIs and 650 TMAs from 321 different patients. A detailed dataset summary can be found in Table 1. This comprehensive dataset facilitates robust survival prediction modeling and validation.

The protocol is based on a multicenter study with a short planned recruitment period, ensuring that patients received similar postoperative care across 14 study centers located in 6 different German states over three years. If PDAC is suspected and curative surgery is performed, the patient can enroll in the study. After successful resection, the pathologist prepares paraffin-embedded tumor samples and samples of adjacent pancreatic tissue from the surgical specimen and sends them to the University Hospital of Cologne. Tissue samples are collected strictly according to TNM (Tumour, Node, Metastasis) staging, and patients with no remaining tumor tissue are excluded from the study protocol. The objective is to study a population that might benefit from surgery and additional biomarker-driven therapy in the future. Clinical data was collected over a 2-year follow-up observation period, with updates every 6 months by the coordinating center.

| LEVEL | STATISTICS | PANCALYZE |
|---|---|---|
| OVERALL | # Patients | 321 |
| | Death ratio | 63% |
| WSI | # Samples | 850 |
| | Patches | 2,816,120 |
| | Sampled patches | 998,553 |
| TMA | # Samples | 650 |
| | Patches | 210,600 |
| | Sampled patches | 162,500 |
| AVG. PER PATIENT | WSI | 3 |
| | TMA | 2 |

Table 1: PANCALYZE Dataset - Study Statistics

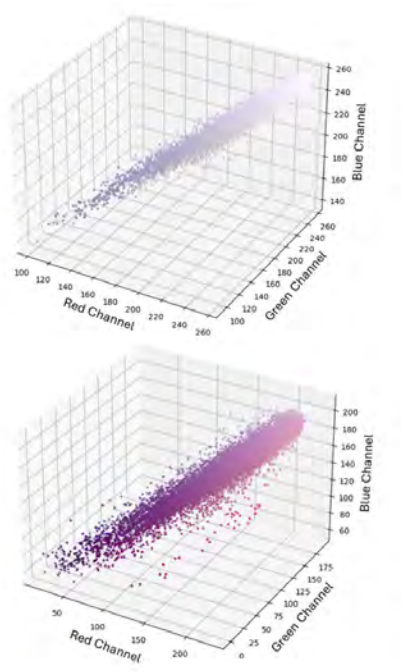



Figure 4: The top figure shows the raw version of the TMA images with pale colors and the bottom part shows result after color correction based on (Reinhard et al., 2001).

### *3.2. GraphLSurv*

The state of the art in computational pathology demonstrates significant advancements in using deep learning and GCNs for survival prediction. The approach of, GraphLSurv (Liu et al., 2023a), introduces a scalable graph convolutional network designed for survival prediction by generating adaptive and sparse graph structures to model the complex relationships between patches.

#### *3.2.1. Preprocessing*

The PANCALYZE dataset includes two distinct histopathological techniques: Tissue Microarrays (TMAs) and Whole Slide Images (WSIs). Initially, tissue sections are obtained and samples are cut, from which whole slide images (WSIs) are created through scanning and digitization. Subsequently, tissue microarrays (TMAs) are constructed as an independent process by extracting cores from these original tissue sections and arraying them together (Pilla et al., 2012). This distinction is crucial as both techniques exhibit different staining intensities. These variations can be attributed to differences in the time of digitization and scanner configurations.

*Color Enhancement.* Due to these differences, the preprocessing pipeline required adjustments for both TMA and WSI samples, since the beginning of the pipeline, as depicted in Figure 4. The initial distribution of the RGB channels in TMAs and WSIs images appeared pale,

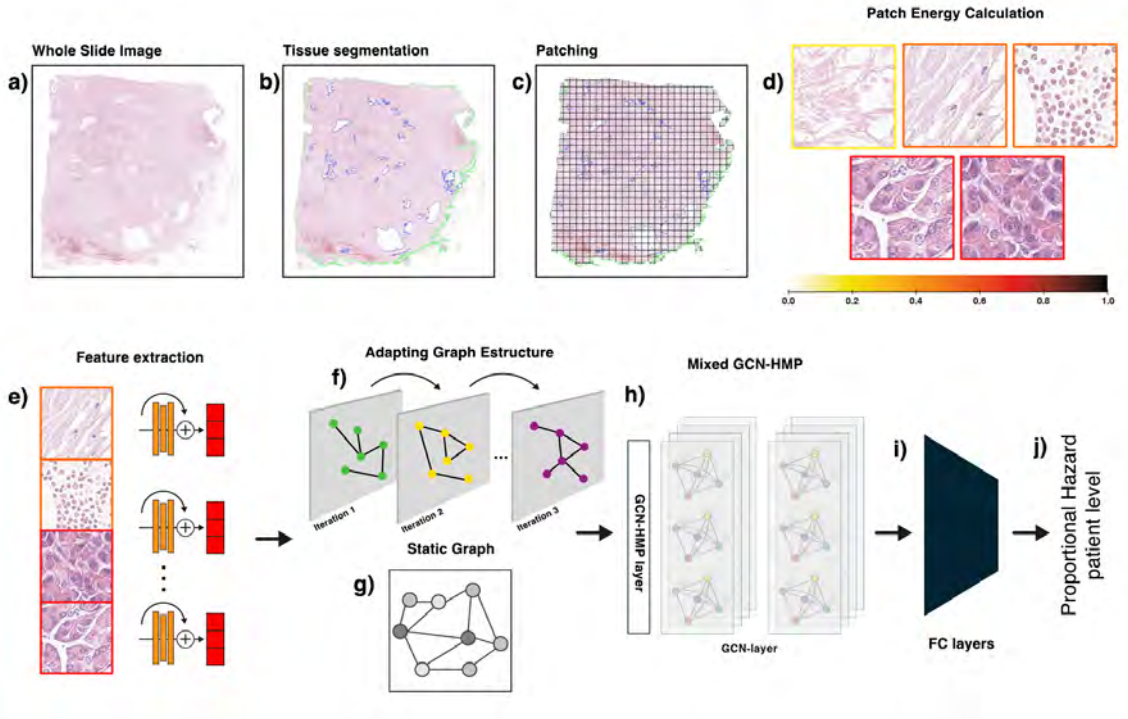



Figure 3: Representation of the pipeline from the raw original image to the computation of the proportional hazard at the patient level. This weakly supervised framework is used for survival prediction on WSIs and TMAs. Non-overlapping 256x256 patches are extracted, and following energy calculation, the top patches are fed into ResNet-50 (CLAM) and ViT/16 via DINOv2 (UNI). The survival-aware structure learning then utilizes these patch features to construct an adaptive and sparse structure.

making it hard to achieve proper segmentation of the tissue. To correct this, a "Color Transfer" technique was implemented (Reinhard et al., 2001). This technique efficiently transfers the color characteristics from one image (source) to another (target). Where in this case a source image was taken from the dataset of TCGA-BRCA (Kandoth et al., 2013), which shows more intense colors in the H&E staining. The color correction involves a statistical analysis of the color distributions in both images, followed by a transformation of the target image to match the color distribution characteristics (mean and variance) of the source image. This method is computationally efficient and suitable for large-scale datasets.

*Tissue Segmentation.* After color correction, each digitized file undergoes a uniform processing pipeline, starting with automated segmentation of tissue sections. The images are converted from RGB to HSV color space and loaded into memory at a downsampled resolution. A binary mask of the foreground tissue areas is created by thresholding the saturation channel. To refine the mask, median blurring is applied to smooth edges, and morphological closing is performed to fill small gaps and holes. The detected contours of the foreground objects are then filtered based on an area threshold and subsequently saved.

Non-overlapping patches of 256x256 pixels are extracted at 20x magnification from the segmented tissue regions. This results in thousands of patches per WSI, capturing detailed histopathological features. Training on all image patches is extremely time-consuming. Using a filtering technique to select relevant patches enhances computational efficiency and improves model performance by focusing on the most informative regions, thereby avoiding redundancy and noise. GCNs excel at capturing spatial relationships and dependencies between different regions within an image, and selecting relevant patches can help the model to better learn these relationships (Adnan et al., 2020), due to this an "Energy Calculation" sampling strategy, is implemented to filter patches lacking obvious texture, by computing the energy map for each patch via an efficient Sobel filter implementation (Avidan and Shamir, 2023). Patches exhibiting pronounced texture typically yield higher energy values, indicative of strong gradients. Conversely, regions with uniform textures or smooth gradients register lower energy values. As shown in Figure 5 depicts a scale ranging from 0 to 1. This scale highlights the patches based on their content relevance. Utilizing these energy maps, the algorithm proceeds to identify the top $s$ patches with the most significant energy values. For Whole Slide Images (WSI), $s$ is set to 1,000, whereas for Tissue Microarrays (TMA), it is 250. Furthermore, adhering to standard preprocessing protocols, applying the classical Color Normalization tech-

nique (Macenko et al., 2009) on all patches to mitigate color disparities among various patches.

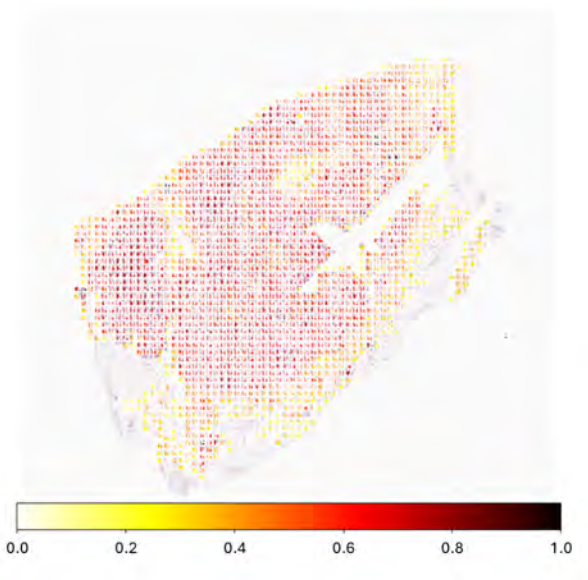



Figure 5: Visualization of energy calculation using the Sobel filter overlaid on a raw Whole Slide Image (WSI). Patches with high energy levels, indicative of their relevance, are accentuated in red or dark shading, while low-energy and less significant patches are represented in yellow or white tones.

*Feature extraction.* In this methodology, we explored two distinct feature extractors, the default extractor proposed in the original experiment is CLAM (Lu et al., 2021), a modified ResNet-50 model pre-trained on ImageNet (Deng et al., 2009). Retaining only the initial convolutional layers (conv1) and subsequent convolutional blocks (conv2_x, conv3_x, conv4_x), CLAM concludes with an average pooling layer (avgpool). This modification allows the process of the original image tile data, represented as (Width, Height, 3 channels), resulting in feature maps with dimensions of (Width/16, Height/16, 1024 channels). Subsequently, an Adaptive Average Pooling layer (AdaptiveAvgPool2d(1)) collapses the spatial dimensions, yielding a final feature vector of length 1024. This process facilitates the extraction of comprehensive low-level features from image tiles, crucial for subsequent analysis and model training.

Additionally, the second feature extractor UNI (Chen et al., 2024), which is a pretrained vision backbone based on the ViT-L/16 (Beyer et al., 2022) model architecture, trained using the DINOv2 (Oquab et al., 2023) self-supervised learning algorithm. Applying the Mass-100K dataset, which comprises over 100 million histology images from diverse sources, UNI is adept at capturing the intricate features and patterns present in WSIs. During inference, images are resized and normalized using ImageNet parameters before being passed through the model to extract features.

Both extracted features have as well a final feature vector of length 1024.

### *3.2.2. Graph Construction*

Survival analyses typically incorporate datasets that include individual attributes (such as demographic details and clinical records), the duration of follow-up $t \in \mathbb{R}$, and the outcome status $\sigma \in \{0, 1\}$. For this research, the data is represented from $n$ patients with the collection:

$$\{(x_i, t_i, \sigma_i)\}_{i=1}^{n} \tag{2}$$

in which $x_i \in \mathbb{R}^d$ signifies the characteristics of the $i$-th patient, $t_i$ indicates the observation period for the $i$-th patient, and $\sigma_i$ reflects the event status at the observation time $t_i$. Here, $\sigma = 0$ signifies no event occurrence, designating the data as right-censored, while $\sigma = 1$ indicates an event occurrence. These right-censored observations are included in our analysis. The data set for each patient is expressed as:

$$B_i = \{I_j \in \mathbb{R}^c\}_{j=1}^{s} \tag{3}$$

where $I_j$ is the embedding for the $j$-th patch. The full dataset is thus defined as:

$$\{(B_i, t_i, \sigma_i)\}_{i=1}^{n} \tag{4}$$

*GraphLSurv structure learning.* The study uses a strategy for learning structures that dynamically generate sparse and adaptive graphs. Using a technique of patch similarity learning, aiming to optimally connect patches to effectively highlight dense and meaningful areas within the structures. Unlike traditional methods such as those used in DeepGraphSurv (Li et al., 2018), this approach directly computes connections without relying on pre-existing computational models.

The input feature matrices of patients, represented as $\{X_i\}_{i=1}^{n}$ are processed, along with a learnable projection matrix $T \in \mathbb{R}^{c \times p}$, and a defined threshold $\delta$. Each feature matrix $X$ is transformed into a new vector space by $T$, resulting in a transformed feature matrix $P \in \mathbb{R}^{s \times p}$. This matrix undergoes pairwise cosine distance evaluation to assess patch similarities, forming a symmetric adjacency matrix $A_L$. Connections in $A_L$ weaker than $\delta$ are nullified to enhance the graph's sparsity, ensuring that only significant and relevant connections are preserved.

This innovative framework encapsulated in the function $M(\cdot)$, defined as:

$$M(X; T, \delta) : \mathbb{R}^{s \times c} \rightarrow \mathbb{R}^{s \times s} \tag{5}$$

transforms the raw feature space into one that is optimally configured for generating adaptive structures. When applied to survival prediction tasks, these graphs evolve to become survival-aware, enhancing their utility as the network is optimized.

For the k-nearest neighbors (k-NN) method, the adjacency matrix, denoted as $A_I \in \mathbb{R}^{s\times s}$, is computed based on the nearest neighbors, discarding edges that exceed a predefined threshold. The radius-based method, in contrast, establishes adjacencies through geographical proximity within a specified radius.

The methodology described not only facilitates the exploration of spatial and feature-based relationships but also significantly advances the analysis of complex image data structures within medical imaging and spatial data fields. The adaptability of the graph construction process allows for more tailored and effective training of graph convolutional networks (GCNs) (Müller et al., 2023), particularly beneficial in computational pathology and related areas of research.

Graph convolutional techniques are applied to update patch features by allowing non-local node embedding learning through feature aggregation across graphs. This advanced method enables each patch to integrate information from connected patches, effectively enriching its feature set with expansive and comprehensive data inputs. The primary k-nearest neighbors (k-NN) graph, $A_I \in \mathbb{R}^{s\times s}$, constructed from raw feature data, plays a critical role in survival prediction. This graph, combined with an adaptive graph $A_L$, forms a hybrid graph where hybrid message passing (HMP) is executed.

The transformation and aggregation process is governed by the function:

$$F(X|A_I, A_L) : \mathbb{R}^{s\times c} \rightarrow \mathbb{R}^{s\times h} \tag{6}$$

Here, $h$ denotes the output dimension, and the feature matrix $X$ is dynamically updated via the equation:

$$F(X) = \lambda X_I + (1-\lambda)X_L \tag{7}$$

The parameter $\lambda$ balances the influence of the initial k-NN graph $A_I$, enhancing the model's ability to adapt based on the evolving graph structure.

Each patient's sample bag feature matrix $X$ is refined through this function, resulting in a new matrix $E \in \mathbb{R}^{s\times h}$. To ensure efficient memory usage, a singular graph convolution layer is employed to process both $A_I$ and $A_L$. Further enhancement of patch features is achieved by applying multiple layers of graph convolutions that implement HMP. The final representations of the WSI and TMA, are designated as $S_{\text{rep}} \in \mathbb{R}^{2h}$, which is derived by combining the results from two prevalent graph pooling operations: maximum and average pooling.

*Dynamic GCN Enhacements* . To optimize traditional Graph Convolutional Networks (GCNs) (Kipf and Welling, 2016), which generally operate on fixed graph structures, the GraphLSurv (Liu et al., 2023b) framework introduces a GCN-HMP layer. This layer facilitates hybrid message passing and adaptive structure learning, allowing dynamic updates to the graph as processing progresses. It functions through an innovative method where a function, $F$, conditions the input features, $X$, on a hybrid graph structure $A_I$ combined with an adaptive mechanism $M(X)$. This approach enables the GCN to not only process but also adaptively learn from the data, enriching the feature interactions dynamically across the graph. This is implemented by:

$$\text{GCN}_{\text{HMP}}(X, A_I) = F(X|A_I, M(X)). \tag{8}$$

Within this framework, two distinct architectures are implemented: pure GCN-HMP and mixed GCN-HMP. The pure version updates its graph structure at every layer to reflect new data relationships, thereby enhancing adaptability. Conversely, the mixed version maintains a static graph structure after its initial computation, thereby providing stability across deeper network layers.

The computational demands associated with large-scale datasets, like WSIs containing thousands of patches, are effectively managed by employing an anchor-based strategy. This strategy utilizes a subset of data, referred to as 'anchors' (Chen et al., 2020), to approximate the complete graph structure, substantially reducing computational complexity from quadratic to linear relative to the number of patches. It calculates cosine distances between patches and their corresponding anchors, followed by focused message passing operations tailored for this graph representation.

*Survival Prediction.* The computed representation of WSI and TMA are utilized to predict the survival risk, $\hat{y}_i \in \mathbb{R}$, of the $i$-th patient through a series of fully connected layers. To ensure a smooth transition of features across connected nodes in the graph, thereby promoting continuity in the structural representation defined by the patch matrix $X$ and the adjacency matrix $A$, Dirichlet energy (Belkin and Niyogi, 2001) is incorporated into the loss function. This inclusion encourages homogeneity in node features, penalizing large variations between connected nodes to enhance the model's ability to generalize across similar structures. This inclusion is mathematically represented by:

$$\ell_{\text{graph}} = \frac{1}{s^2}\,\text{tr}(X^T L X) \tag{9}$$

where $L = D - A$ denotes the graph Laplacian, $D$ is the degree matrix with $D = \sum_j A_{i,j}$, and $\text{tr}(\cdot)$ is the trace matrix. This term facilitates the smooth variation of patch embeddings across adjacent patches, enhancing the model's ability to capture subtle nuances in data relationships. Consequently, the overall loss function is formulated as:

$$\ell_{=}\ell_{\text{cox}} + \alpha\ell_{\text{graph}} \tag{10}$$

where $\alpha$ is a hyper-parameter within the range [0, 1]. This parameter balances the influence of the structural

smoothness on the model's performance, allowing for the adjustment of the regularization strength based on the specific analytical needs.

### *3.3. Proposed Experiments*

Two preprocessing tools were employed in the experimental setup to harness diverse computational backbones and test the robustness of the GraphLSurv model across different image modalities. The first tool, CLAM (Lu et al., 2021), utilizes a ResNet-50 backbone to preprocess the PANCALYZE dataset, which includes both Whole Slide Images (WSIs) and Tissue Microarrays (TMAs). This setup enables the application of GraphLSurv pipeline for graph construction and subsequent survival prediction.

In a parallel experiment, the preprocessing tool UNI (Chen et al., 2024), which harnesses the capabilities of a Vision Transformer (ViT/16) trained via DINOv2, was used. This approach seeks to exploit the self-attention mechanisms inherent in transformers to better capture the contextual relationships within WSIs and TMAs. Utilizing the same PANCALYZE dataset, this method extends the scope of GraphLSurv to examine the impact of advanced image features extracted via state-of-the-art unsupervised learning techniques on the efficacy of graph-based survival prediction models.

*Implementation Protocol.* To ensure robust model evaluation the dataset was initially partitioned with 80% allocated to training and 20% to testing, ensuring that each fold was representative of the overall dataset. Additionally, 20% of the training subset was reserved for validation purposes, facilitating model tuning and early stopping without overfitting.

The stratification was conducted at the patient level, adhering to the proportions of censored cases as per the methodologies established in WSISA (Zhu et al., 2017). This stratification ensures that each fold maintains a balanced representation of the survival outcomes, crucial for training survival prediction models that are sensitive to outcome distributions. The efficacy of the data splitting strategy was validated using the log-rank test, yielding a p-value of approximately 0.9. This statistic confirms the absence of significant differences in survival distributions among the training, validation, and testing sets, thereby supporting the integrity of the experimental design.

Experiments were executed under defined random conditions (42) to ensure repeatability and reliability in statistical results. To construct graph structures, different k-nearest neighbor (k-NN) configurations [4, 6, 8, 10, 12] were utilized to explore various levels of connectivity. These graphs were designed with dataset splits performed without stratification, and with a mechanism to log predictions for detailed outcome analysis.

*Training Configuration.* Training was executed in batches of one, optimizing parallel data processing with eight workers. Optimization was driven by an Adam optimizer, configured with an initialization learning rate of $10^{-4}$ and adjusted dynamically based on performance plateaus lr_factor: 0.8, lr_patience: 10, lr_min: $10^{-5}$. The inclusion of weight decay added a regularization effect, crucial for managing complexity and overfitting in deep learning networks. The training span covered 300 epochs, incorporating an early stopping mechanism activated after 30 epochs without improvement (patience: 30), enhancing computational efficiency and model performance.

*Graph Learning Regularization.* Graph regularization was strategically enabled to promote model generalization. This feature incorporated three distinct regularization strategies: smoothness ($\alpha$), degree, and sparsity. These parameters are essential for refining the learning dynamics within the graph, affecting how features are propagated and influencing overall model stability and interoperability.

*Anchor-Based Enhancements.* The anchor graph learner was defined with a hidden dimension of 128, utilizing 20% of data points as anchors, essential for optimizing the graph structure. The method uses a transformer-based metric and controls the threshold for anchor connections with a parameter ($\delta = 0.8$), which determines the sensitivity of the model to the proximity of nodes. The anchor graph encoder layer operates through a single graph hop, with the initial graph contribution set at 50% ($\lambda = 0.5$), balancing adaptability and structural consistency in graph processing.

## 4. Results

This section presents the findings obtained from applying GraphLSurv to a PANCALYZE dataset of Pancreatic Ductal Adenocarcinoma (PDAC) using two histopathology modalities, Whole Slide Images (WSI) and Tissue Microarrays (TMA), with two different feature extractors: ResNet-50 by CLAM (Lu et al., 2021) and ViT/16 via DINOv2 by UNI (Chen et al., 2024). The aim was to assess the performance of survival prediction across these modalities and feature extractors. The results are summarized as follows:

### *4.1. Concordance Index (C-Index)*

The Concordance Index (C-Index) serves as a fundamental metric for evaluating the efficacy of survival prediction models. In addition to choosing the feature extractor, we performed parameter selection based on outcomes from the validation set. This extended to investigating how different structural configurations impact model performance. Our analysis revealed significant variability in the C-Index for fixed k-NN structures

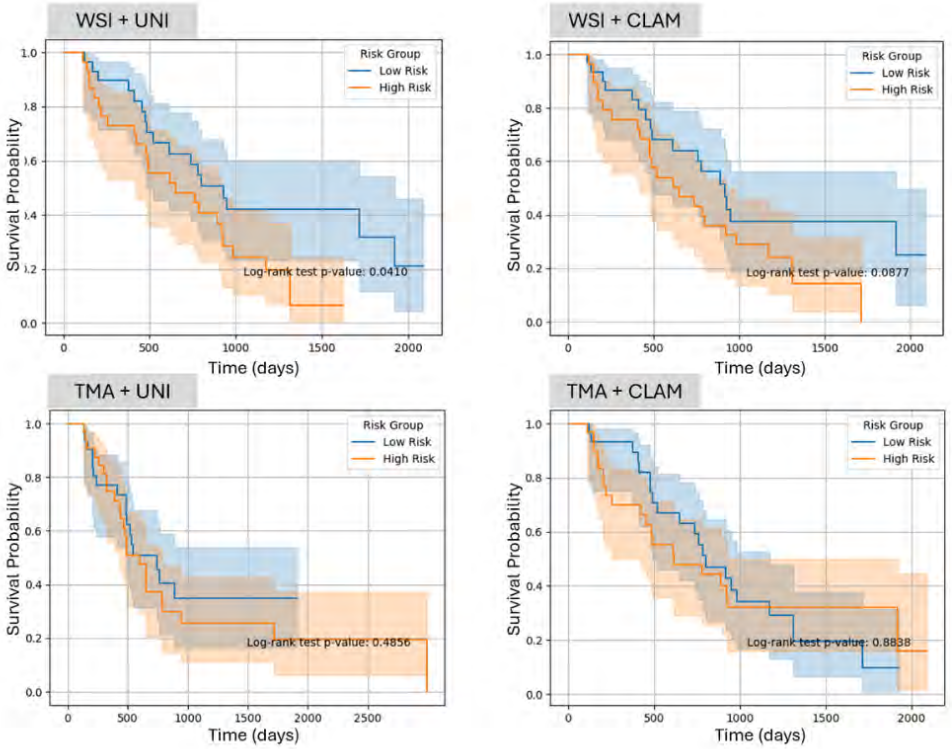



Figure 6: Kaplan-Meier curves showing patient stratification, with high and low-risk groups compared using the log-rank test p-value. The figure contrasts histopathology modalities: TMAs and WSIs, each analyzed with UNI and CLAM feature extractors.

| | UNI | CLAM |
|---|---|---|
| k | C-index | C-index |
| 4 | 0.5925 | 0.4943 |
| 6 | 0.5320 | 0.5378 |
| 8 | 0.5872 | 0.4975 |
| 10 | 0.5968 | 0.5228 |
| 12 | 0.5984 | 0.5233 |

Table 2: Comparison of C-index for UNI (ViT/16 via DINOv2) and CLAM (ResNet-50) across different values of k for WSIs.

across various 'k' values, underscoring its sensitivity to fine-tuning needs.

In our quest to determine the optimal settings for each feature extractor, we conducted experiments with varying values of 'k' (4, 6, 8, 10, 12) as shown in the Table 2. Thorough comparative analysis between CLAM (ResNet-50) and UNI (ViT/16 via DINOv2), we gained valuable insights into the stability and efficacy of different configurations. For instance, the C-Index for WSI + CLAM was recorded at 0.5378 using (k = 6). However, this decreased to 0.4970 for TMA + CLAM. Conversely, WSI + UNI with (k = 12) improved to 0.5984, whereas TMA + UNI yielded a lower score of 0.5203.

Our exploration into the impact of structure sparsity on model performance provided valuable insights into optimizing GraphLSurv. By adjusting the parameters to explore sparser structures, we observed a discernible correlation between high sparsity and improved model performance in terms of the C-Index. However, it's essential to exercise caution, as excessively sparse or dense structures may not yield optimal results. Additionally, our study delved into the optimal utilization of structure learning within the mixed GCN-HMP model, where through experiments and fine-tuning the validation results indicated that a single application of structure learning sufficed for effective survival prediction, suggesting that incorporating structure learning at multiple layers might introduce unnecessary complexity without significant performance gains.

### *4.2. Prognostic Risk Group Analysis*

An essential aspect of evaluating survival prediction models lies in their capacity to stratify patients into distinct risk categories based on prognostic outcomes. In our study, this stratification was achieved through the application of Kaplan-Meier curves, complemented by the log-rank test to ascertain statistical significance via p-values, which in this study are evaluated from the test set.

In the conducted experiments, differences in survival

outcomes between high-risk and low-risk groups were observed for only one feature extractor, as shown in Figure 6. For WSI + CLAM, an interesting result is presented (p-value = 0.0877), but this cannot be considered statistically relevant, as it exceeds the threshold p-value of $\leq 0.05$. For TMA + CLAM, results show there is no ability to differentiate groups (p-value = 0.8838). On the other hand, in the UNI experiment, distinctions were shown for WSI + UNI (p-value = 0.0410), but not with the same success for TMA + UNI (p-value = 0.4856). These findings underscore the predictive capability of the models across various modalities and feature extractors

Examining the Kaplan-Meier curves shows a clear difference between the two feature extractors, CLAM (ResNet-50) and UNI (ViT/16 via DINOv2), in their performance with TMA and WSI modalities. WSI + UNI demonstrates better ability to distinguish between risk groups, as indicated by significant differentiation and a lower p-value.

In the TMA modality, both feature extractors struggle to clearly distinguish between high-risk and low-risk groups. While WSI + CLAM shows significant differentiation, there is still room for improvement in consistency between validation and test results. The Kaplan-Meier graphs for TMA + CLAM and TMA + UNI show less distinct separation between risk groups, indicating difficulty in differentiation.

### *4.3. Heatmap Visualization*

Heatmap visualizations were employed to enhance the interpretability of the trained weakly-supervised deep learning classifier. By identifying and aggregating regions within Whole Slide Images (WSIs) with high diagnostic importance, as indicated by elevated attention scores, the classifier facilitated the recognition of morphological features crucial for clinical diagnosis. This process involved disregarding regions considered to have low diagnostic relevance, thereby focusing attention on areas pivotal for accurate prediction. The resultant heatmaps, derived from the model's attention scores, provided insights into the relative significance of each region within the WSIs, helping in the delineation of boundaries between tumor and normal tissue. As shown in the Figure 7, a comparison between censored and deceased tissues was conducted, along with an analysis of the highest attention score patches. In the censored section, this comparison revealed distinct patterns in the content of patches where the highest attention was retained, differing from those selected for deceased tissues.

## 5. Discussion

Survival prediction for pancreatic cancer using deep learning techniques remains an active yet challenging area of research, primarily due to the scarcity of large annotated datasets. The extremely low prevalence of pancreatic cancer complicates the acquisition of extensive datasets with accurate survival annotations, which is critical for developing robust and generalizable models. Annotating medical images and clinical data for pancreatic cancer is labor-intensive and requires expert knowledge, creating a significant bottleneck that slows the creation of large annotated datasets necessary for effective model training. Additionally, the lack of standardized evaluation protocols for deep learning models in survival prediction complicates the comparison of different approaches and their applicability in real-world settings.

GraphLSurv, a graph-based survival analysis model, was evaluated with two distinct feature extractors: ResNet-50 by CLAM and ViT-L/16 via DINOv2 by UNI. The results affirm the feasibility of graph-based methods for survival prediction, with GraphLSurv effectively learning adaptive and sparse structures to capture essential correlations between image patches. The model demonstrated good results using the proposed adaptive fixed k-NN structures, which exhibited significant variability. This underscores the adaptive structure's enhanced capability to capture critical correlations necessary for precise survival prediction. The hybrid structure ($\lambda = 0.5$) showed promising results, suggesting that a balanced approach might help improve the predictive c-index.

Kaplan-Meier curves and log-rank tests validated the model's efficacy in stratifying patients into distinct risk categories, with the WSI + UNI configuration exhibiting superior discriminative ability. The statistical significance of these stratifications highlights the model's potential clinical utility in identifying high-risk patients who may benefit from more aggressive treatment regimens.

Despite these promising results, our study has several limitations. A primary challenge lies in learning sparse structures, necessitating meticulous threshold adjustments. Moreover, there is a need to improve the transparency of graph construction and attention weight mechanisms to enhance the model's clinical utility. Addressing these issues could lead to more interpretable and reliable models, making them more acceptable for clinical use.

Another limitation is the absence of a model trained exclusively on pancreatic cancer data. Although UNI was trained on histopathology images, pancreatic cancer remains underrepresented in the dataset. This underscores the need for extensive, disease-specific datasets to bolster model performance. UNI, trained on histopathology images, ranks among the top five least represented cancer types, with fewer than 4,000 slides, whereas models for heart, lung, and kidney cancers were trained on approximately 10,000 slides. This disparity suggests that increasing the representation of

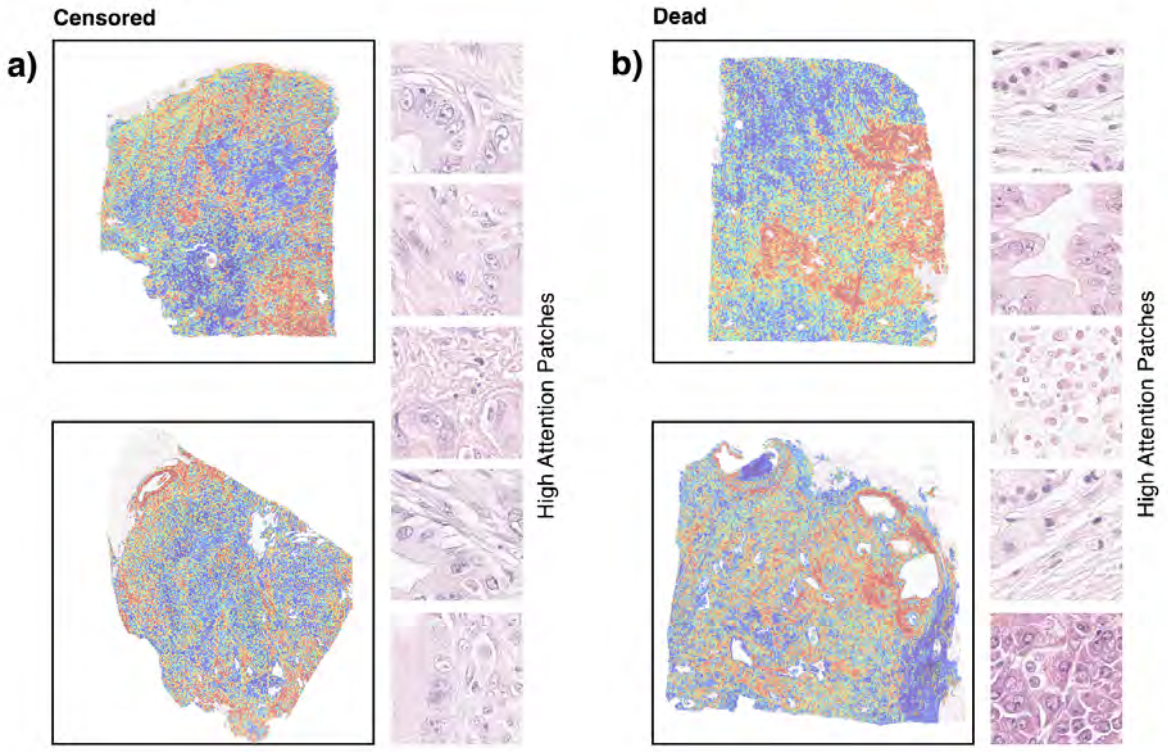



Figure 7: The figure shows WSI visualizations and top 5 high-attention patches for censored patients (left) and dead patients (right), with whole slide attention heatmaps where red indicates high attention and blue indicates low attention.

pancreatic cancer in training datasets could significantly enhance model accuracy and generalizability.

However, the dataset's limitations are notable. The data was acquired solely from institutions in Germany, which employ relatively uniform and standardized tissue processing and staining protocols. This uniformity may limit the generalizability of the model across diverse clinical settings with varying protocols. Future research should focus on developing models using data from multiple institutions with diverse tissue processing and staining protocols to enhance model robustness and generalizability across various clinical settings.

It was also noted that the CLAM feature extractor, trained on ImageNet, was less effective than UNI, which was trained on specific histopathology images. This suggests that feature extractors trained on domain-specific data may offer superior performance in medical image analysis tasks. Additionally, WSI + CLAM's performance was inconsistent during training, indicating difficulty in distinguishing between groups.

Incorporating manual annotations, such as detailed tissue region and cell annotations, could further improve performance and interpretability. These annotations would provide more granular information, enabling the model to learn finer distinctions in tissue morphology that are critical for accurate survival predictions. Additionally, incorporating biomarkers and clinical data could provide valuable features not captured by images alone, enhancing the model's decision-making process. For instance, biomarkers can help identify patients at higher risk of metastasis, offering another layer of prognostic information that can be integrated into the model. Exploring TMAs with various staining techniques or biomarkers could offer a comprehensive approach, providing additional information for deep learning models. This combined strategy has the potential to improve survival prediction accuracy.

This study addressed the complexities of predicting pancreatic cancer survival using Whole Slide Images (WSIs) and Tissue Microarrays (TMAs). WSIs include vast amounts of data, demanding significant computational resources, whereas TMAs offer limited data, potentially skewing models towards specific regions of interest (ROIs) and ignoring broader pathological contexts. Additionally, the small size of TMA patches ( 300 patches, 256x256) did not provide sufficient information for the network to make accurate decisions, which may have contributed to the inferior performance observed in TMA-based analyses.

## 6. Conclusions

This study demonstrates the potential of graph-based survival prediction models for PDAC. The findings

suggest that adaptive and sparse structures outperform static or dense ones for modeling WSIs and TMAs. Nonetheless, further research is necessary to refine these models, enhance their explainability, and ensure their applicability across diverse clinical environments. Advancements in this domain could significantly impact clinical decision-making and improve outcomes for PDAC patients. By addressing the limitations and incorporating more diverse and comprehensive datasets, future studies can build on these findings to develop even more robust and clinically useful survival prediction models.

## Acknowledgments

I deeply want to thank my supervisors for their guidance, teaching, trust, and for the opportunity to have developed this project under their institution. I also want to express my gratitude for the support I received from my family, who always believe in my projects and help me achieve them regardless of the complications along the way. To my friends who, regardless of the time or distance, were always there, motivating me and helping me in whatever was needed.

## References

Adnan, M., Kalra, S., Tizhoosh, H.R., 2020. Representation learning of histopathology images using graph neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 988–989.

Al-Janabi, S., Huisman, A., Diest, P.J.V., 2012. Digital pathology: current status and future perspectives. Histopathology 61, 1–9. doi:10.1111/j.1365-2559.2011.03814.x.

Avidan, S., Shamir, A., 2023. Seam carving for content-aware image resizing, in: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 609–617.

Baek, E.T., Yang, H.J., Kim, S.H., Lee, G.S., Oh, I.J., Kang, S.R., Min, J.J., 2021. Survival time prediction by integrating cox proportional hazards network and distribution function network. BMC bioinformatics 22, 1–15.

Belkin, M., Niyogi, P., 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. Advances in neural information processing systems 14.

Beyer, L., Zhai, X., Kolesnikov, A., 2022. Big vision. https://github.com/google-research/big_vision. GitHub repository.

Bradburn, M.J., Clark, T.G., Love, S.B., Altman, D.G., 2003. Survival analysis part ii: multivariate data analysis–an introduction to concepts and methods. British journal of cancer 89, 431–436.

Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A.H., Shaban, M., et al., 2024. Towards a general-purpose foundation model for computational pathology. Nature Medicine .

Chen, R.J., Lu, M.Y., Shaban, M., Chen, C., Chen, T.Y., Williamson, D.F., Mahmood, F., 2021. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference, Strasbourg, France, September 27 – October 1, 2021, Proceedings, Part VIII, Springer. pp. 339–349. doi:10.1007/978-3-030-87237-3_33.

Chen, Y., Wu, L., Zaki, M., 2020. Iterative deep graph learning for graph neural networks: Better and robust node embeddings. Advances in neural information processing systems 33, 19314–19326.

Clark, T.G., Bradburn, M.J., Love, S.B., Altman, D.G., 2003. Survival analysis part i: basic concepts and first analyses. British journal of cancer 89, 232–238.

Cooper, M., Ji, Z., Krishnan, R.G., 2023. Machine learning in computational histopathology: Challenges and opportunities. Genes, Chromosomes and Cancer 62, 540–556. doi:10.1002/gcc.23177.

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.

Hidalgo, M., Cascinu, S., Kleeff, J., Labianca, R., Löhr, J.M., Neoptolemos, J., Real, F.X., Van Laethem, J.L., Heinemann, V., 2015. Addressing the challenges of pancreatic cancer: future directions for improving outcomes. Pancreatology 15, 8–18.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al., 2013. Mutational landscape and significance across 12 major cancer types. Nature 502, 333–339.

Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y., 2018. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. BMC medical research methodology 18, 1–12.

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 .

Li, G., Chen, J.Z., Chen, S., Lin, S.Z., Pan, W., Meng, Z.W., Cai, X.R., Chen, Y.L., 2020. Development and validation of novel nomograms for predicting the survival of patients after surgical resection of pancreatic ductal adenocarcinoma. Cancer medicine 9, 3353–3370.

Li, K., Qian, Z., Han, Y., Eric, I., Chang, C., Wei, B., Lai, M., Liao, J., Fan, Y., Xu, Y., 2023. Weakly supervised histopathology image segmentation with self-attention. Medical Image Analysis 86, 102791.

Li, R., Yao, J., Zhu, X., Li, Y., Huang, J., 2018. Graph cnn for survival analysis on whole slide pathological images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 174–182.

Liu, P., Ji, L., Ye, F., Fu, B., 2023a. Graphlsurv: A scalable survival prediction network with adaptive and sparse structure learning for histopathological whole-slide images. Computer Methods and Programs in Biomedicine 231, 107433.

Liu, P., Ji, L., Ye, F., Fu, B., 2023b. GraphLSurv: A scalable survival prediction network with adaptive and sparse structure learning for histopathological whole-slide images. Computer Methods and Programs in Biomedicine 231, 107433. doi:doi.org/10.1016/j.cmpb.2023.107433.

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.

Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E., 2009. A method for normalizing histology slides for quantitative analysis, in: 2009 IEEE international symposium on biomedical imaging: from nano to macro, IEEE. pp. 1107–1110.

Mizrahi, J.D., Surana, R., Valle, J.W., Shroff, R.T., 2020. Pancreatic cancer. The Lancet 395, 2008–2020.

Müller, T.T., Starck, S., Dima, A., Wunderlich, S., Bintsi, K.M., Zaripova, K., Braren, R., Rueckert, D., Kazi, A., Kaissis, G., 2023. A survey on graph construction for geometric deep learning in medicine: Methods and recommendations. Transactions on Machine Learning Research .

Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P., 2023. Dinov2: Learning robust visual features without supervision.

Pilla, D., Bosisio, F.M., Marotta, R., Faggi, S., Forlani, P., Falavigna, M., Biunno, I., Martella, E., De Blasio, P., Borghesi, S., et al.,

2012. Tissue microarray design and construction for scientific, industrial and diagnostic use. Journal of Pathology Informatics 3, 42.

Popp, F.C., Popp, M.C., Zhao, Y., Betzler, C., Kropf, S., Garlipp, B., Benckert, C., Kalinski, T., Lippert, H., Bruns, C.J., 2017. Protocol of the pancalyze trial: a multicenter, prospective study investigating the tumor biomarkers cxcr4, smad4, sox9 and ifit3 in patients with resected pancreatic adenocarcinoma to predict the pattern of recurrence of the disease. BMC cancer 17, 1–9.

Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P., 2001. Color transfer between images. IEEE Computer Graphics and Applications 21, 34–41.

Ren, M., Zhang, Q., Zhang, S., Zhong, T., Huang, J., Ma, S., 2022. Hierarchical cancer heterogeneity analysis based on histopathological imaging features. Biometrics 78, 1579–1591.

Rossing, H.H., Talman, M.L.M., Lænkholm, A.V., Wielenga, V.T., 2012. Implementation of tma and digitalization in routine diagnostics of breast pathology. Apmis 120, 341–347. doi:10.1111/j.1600-0463.2011.02871.x.

Sandarenu, P., Millar, E.K., Song, Y., Browne, L., Beretov, J., Lynch, J., Graham, P.H., Jonnagaddala, J., Hawkins, N., Huang, J., et al., 2022. Survival prediction in triple negative breast cancer using multiple instance learning of histopathological images. Scientific Reports 12, 14527.

Wang, Y., Wang, Y.G., Hu, C., Li, M., Fan, Y., Otter, N., Sam, I., Gou, H., Hu, Y., Kwok, T., et al., 2022. Cell graph neural networks enable the precise prediction of patient survival in gastric cancer. NPJ precision oncology 6, 45.

Xu, Z., Lim, S., Shin, H.K., Uhm, K.H., Lu, Y., Jung, S.W., Ko, S.J., 2022. Risk-aware survival time prediction from whole slide pathological images. Scientific reports 12, 21948.

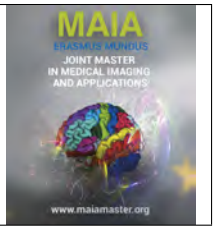
Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., Huang, J., 2020. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Medical Image Analysis 65, 101789.

Zhu, X., Yao, J., Huang, J., 2016. Deep convolutional neural network for survival analysis with pathological images, in: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE. pp. 544–547.

Zhu, X., Yao, J., Zhu, F., Huang, J., 2017. Wsisa: Making survival prediction from whole slide histopathological images, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7234–7242.

Medical Imaging and Applications

Master Thesis, June 2024




# Brain age estimation from MRI images

Clara Lisazo, Adrià Casamitjana, Arnau Oliver and Xavier Lladó

*VICOROB Research Institute, Girona, Spain*

**Abstract**

The accurate estimation of brain age from magnetic resonance imaging scans has significant potential for predicting neurodegenerative diseases and cognitive decline. This study aims to develop and evaluate various pipelines for accurate and robust brain age prediction. We explored both traditional machine learning and deep learning approaches, addressing challenges related to data imbalance and model generalizability. For traditional machine learning models, volumetric measures from cortical and subcortical structures were employed as input features. Deep learning approaches were also explored utilizing diverse input modalities, including T1-weighted (T1w) images, voxel-based morphometry (VBM) volumes, and brain tissue segmentations. Traditional machine learning models demonstrated that volumetric measures from cortical and subcortical structures are valuable predictors of brain age, achieving a mean absolute error (MAE) of 4.18 years on a benchmark dataset (OpenBHB). Deep learning models, particularly the simple fully convolutional network with a soft classification approach, outperformed traditional models. To address age distribution imbalance, strategies such as oversampling and decoupling feature representation training from classifier training were employed, resulting in improved performance for under-represented age groups. The best results were achieved using gray matter VBM volumes, which provided significant information for age differentiation and improved generalization properties. Ensemble learning further enhanced model performance, achieving the best overall MAE of 2.70 years, which outperformed individual models. Additionally, transfer learning from models pretrained on larger datasets significantly improved performance on an in-house dataset, achieving a MAE of 3.28 years. The findings of this work highlight the potential of deep learning methods in developing accurate and generalizable brain age estimation models.

*Keywords:* Brain age estimation, MRI, machine learning, deep learning

## 1. Introduction

The human brain undergoes complex morphological alterations as it ages, including a general decline in volume, cortical thickness and white matter integrity (Driscoll et al., 2009; Hepp et al., 2021; Madden et al., 2009). Changes such as synaptic pruning, demyelination and neurodegenerative processes peak in older age, leading to structural changes like ventricle expansion and cortical thinning (Levakov et al., 2020).

While chronological age serves as an indicator of disease susceptibility, the process and pace of aging differs among individuals, organs, tissues, and medical conditions. As such, there is a growing interest in estimating biological age to better understand aging's diverse manifestations and to predict morbidity and disease. Brain age, estimated from brain magnetic resonance imaging (MRI) scans, is correlated with the biological age and can differ from the chronological age (Hepp et al., 2021). In healthy adults without clinical symptoms, the biological age is anticipated to align with chronological age, on average (Yin et al., 2023).

Positive deviations between brain age and chronological age, known as brain age delta, suggest accelerated brain aging and may indicate underlying pathological processes and disorders (Hepp et al., 2021; Leonardsen et al., 2022; Levakov et al., 2020; Peng et al., 2021). These deviations have been linked to various neurodegenerative diseases, such as Alzheimer's Disease (Gaser et al., 2013; Leonardsen et al., 2022; Salih et al., 2021), depression (Koutsouleris et al., 2014), schizophrenia (Leonardsen et al., 2022; Nenadić et al., 2017), and

multiple sclerosis (Leonardsen et al., 2022), as well as mortality risk (Cole et al., 2018). Therefore, accurate estimation of brain age is crucial for disease risk prediction and early detection of neurodegenerative conditions. Improving brain age estimates can lead to a more accurate detection of deviations from typical aging, and in this way become affordable and noninvasive preclinical biomarkers of early-stage neurodegeneration and cognitive decline (Bellantuono et al., 2021; Yin et al., 2023).

The brain age is often determined through statistical learning techniques using brain scans. Early models were quite simple, usually based on independent voxels or a small number of imaging-derived phenotypes, like volumetric measurements of different anatomical structures or regions of the brain that could reflect brain properties (Franke et al., 2010; Smith et al., 2019). The main limitations of these models were that they were trained on small datasets and generally had limited accuracy and generalization properties. Together with computational advances, the increase in MRI data availability has enabled the development of large-scale deep learning models for precise brain age prediction. Deep learning models can accept minimally or non-preprocessed 3D images as input and model complex nonlinear relationships between voxels, enhancing the accuracy of the age estimations (Leonardsen et al., 2022).

However, the prediction of brain age remains a challenging task, marked by high variability in reported results across studies due to differences in datasets. Factors such as image quality and the age distributions of training data significantly influence the performance of brain age estimation models. Additionally, reliance on data from a single-source dataset, whether public or private, can introduce bias based on the acquisition domain of the images, ultimately limiting the generalization capabilities of the developed models.

To address some of the challenges of brain age estimation from MRI images, this work focuses on developing and evaluating different methods for accurate and robust brain age prediction. We achieve this by training models with a large-scale, multi-site dataset (OpenBHB), which aggregates data from 10 different public sources and comprises 3984 MRI scans originating from 64 different acquisition sites. Utilizing this public dataset facilitates objective evaluation of the different models' performance and enhance their generalizability across different acquisition domains, overcoming limitations associated with single-source datasets. Furthermore, we explore both machine learning and deep learning approaches for the estimation of brain age, conducting a comparative analysis to assess the strengths and weaknesses of each approach in this context. Additionally, we investigate the influence of incorporating prior information, such as brain tissue segmentation, on the performance of deep learning models. This is done to determine if these priors can further improve the accuracy or the robustness of the models. Finally, we study transfer learning techniques to adapt the developed deep learning models, as well as a publicly available pre-trained model, to an in-house dataset. This allows us to identify which approaches achieve superior performance in a real-world clinical setting.

## 2. State of the art

In this section, a review will be made on the current state-of-the-art in brain age estimation, focusing on traditional machine learning and deep learning techniques.

### *2.1. Machine learning approaches*

Machine learning approaches for brain age estimation typically rely on supervised learning techniques, particularly regression analysis. These models treat chronological age as the dependent variable and anatomical brain characteristics extracted from MRI scans as the independent ones. A variety of algorithms have been explored in this domain, including gaussian process regression, Support Vector Regression (SVR), ridge regression, relevance vector regression, XGBoost and ElasticNet (Table 1). This table summarizes studies that employed these machine learning approaches, both classical and more recent ones, for brain age estimation with their respective results. It also includes studies that directly compared the performance of different algorithms on the same datasets (Da Costa et al., 2020; More et al., 2023). While the complexity and computational demands of these models may vary, their performance tends to be comparable (Niu et al., 2020).

Despite the ability of many machine learning algorithms to achieve relatively accurate age estimates based on brain features, it can be observed that there is significant heterogeneity in reported performance across studies. Factors like training and test set size, age range, and feature selection can significantly influence reported model performance measures. Therefore, standardization of methodologies and data is crucial to ensure consistent and reliable results (Soumya Kumari and Sundarrajan, 2024).

### *2.2. Deep learning approaches*

Deep learning models have lately emerged as powerful tools for the estimation of brain age from MRI scans due to their ability to handle 3D MRI data and automatically extract relevant features.

A general overview of some of the latest deep learning methods for brain age estimation is shown in Table 2. This table summarizes, together with the type of architecture, the Mean Absolute Error (MAE) obtained and the datasets used for training, some keypoints of each method. As it can be seen, deep learning architectures for brain age estimation predominantly rely on convolutional neural networks (CNNs), with

Table 1: Summary of machine learning approaches for brain age estimation (SVR: Support Vector Regression, RVR: Relevance Vector Regression, GPR: Gaussian Process Regression).

| Reference | Dataset | Model | MAE |
|---|---|---|---|
| Franke et al. (2010) | IXI, ADNI and private dataset ($N$ = 989, ages = 56.04 ± 11.33) | SVR | 4.98 |
| Mwangi et al. (2013) | INDI dataset ($N$ = 207, ages = 34.9 ± 20.07) | RVR | 5.17 |
| Khundrakpam et al. (2015) | Pediatric MRI Data Repository ($N$ = 308; ages = 12.9 ± 3.8) | ElasticNet | 1.7 |
| Cole et al. (2017) | Brain-Age Healthy Control (BAHC) ($N$ = 2001, ages = 36.95 ± 18.12) | GPR | 4.66 |
| Da Costa et al. (2020) | PAC 2019 ($N$ = 2640, ages = 35.87 ± 16.2) | Linear regression | 13.6 |
| | | SVR | 4.571 |
| | | GPR | 6.13 |
| De Lange et al. (2022) | UK Biobank ($N$ = 41285, ages = 64.15 ± 7.54), CamCAN ($N$ = 622, ages = 54.17 ± 18.4) | XGBoost | 4.18 in UKB; 6.8 in CamCAN |
| More et al. (2023) | CamCAN, IXI, eNKI, 1000BRAINS, CoRR, OASIS-3, MyConnectome, ADNI ($N$ = 2953, ages = 53.27 ± 16.48) | GPR | 4.8 |
| | | RVR | 5.81 |
| | | Ridge regression | 5.64 |
| | | Linear regression | 6.28 |

some models drawing inspiration from the VGG16 architecture (Dinsdale et al., 2021a; Peng et al., 2021). Most studies emphasize the importance of simplicity in the architecture design, as models with fewer parameters tend to yield better results in brain age estimation (Cole et al., 2017; Gong et al., 2021; Peng et al., 2021; Soumya Kumari and Sundarrajan, 2024). The Simple Fully Convolutional Network (SFCN) developed by Peng et al. (2021) exemplifies the success of lightweight models. This architecture, with 7 convolutional blocks (3M parameters), achieves state-of-the-art performance on the UK Biobank dataset (Sudlow et al., 2015) and Predictive Analytics Competition (PAC) 2019 challenge, with MAEs of 2.14 and 3.69 respectively. The lightweight design of this architecture makes it compatible with smaller dataset sizes and 3D volume data. Further demonstrating the effectiveness of this architecture, Leonardsen et al. (2022) implemented variations of SFCN for brain age prediction. They explored three configurations: the soft classification model originally proposed by Peng et al. (2021) (SFCN-sm), a regression variant with a single output neuron (SFCN-reg), and a ranking model (SFCN-rank).

The work of He et al. (2022) stands out as one of the few that incorporates a transformer architecture for brain age estimation. In particular, they employed a dual-pathway architecture that combines a global pathway, trained on the whole 3D MRI volume, for capturing overall brain structure, and a local pathway, trained with smaller 3D patches, for focusing on finer details. An attention mechanism then optimally fuses this global context information with the local details.

Also worth mentioning is that there has been increasing interest in developing interpretable artificial intelligence (AI) models for brain age estimation, since this builds trust in the models' predictions, which is crucial in healthcare applications. Some of the recent studies in this area have focused on achieving explainability, such as the ones of Levakov et al. (2020), Hepp et al. (2021), Wood et al. (2022) and Yin et al. (2023). These attempt to provide spatial maps that highlight the brain areas contributing to predictions.

Furthermore, the study of Hepp et al. (2021) not only focused on interpretability, but also aimed to quantify the model's uncertainty. They achieved this by employing a heteroscedastic noise model. Traditional noise models assume a constant level of noise (error) in the predictions, regardless of the input data. However, in brain age estimation, chronological age is not perfectly encoded within MRI scans. Physiological variations between individuals can introduce inherent uncertainty (aleatoric uncertainty) into the age labels themselves. This means that even for MRI scans that appear very similar visually, the actual chronological ages may differ. A heteroscedastic noise model addresses this challenge by allowing both the mean and variance of the noise to be estimated and vary depending on the specific input data. By incorporating this type of model,

Table 2: Summary of Deep Learning approaches for brain age estimation ('**' denotes the state-of-the-art method on the UK Biobank dataset, whereas '*' denotes the state-of-the-art method on the OpenBHB dataset).

| Reference | Architecture | Dataset | MAE | Keypoints |
|---|---|---|---|---|
| Peng et al. (2021) | 3D CNN (SFCN) | UK Biobank ($N$ = 14503, ages=44-80) | 2.14** | Lightweight fully convolutional network. Treat the regression as a multi-class classification problem. |
| Dinsdale et al. (2021a) | 3D CNN (inspired on VGG) | UK Biobank ($N$ = 12802, ages=44-80) | 2.975 | Evaluated correlations of age predictions with subjects' phenotypical data. |
| Gong et al. (2021) | 3D CNN (SFCN) | PAC 2019 ($N$ = 2638, ages=17-90) | 2.95 | Application of SFCN network to PAC 2019 challenge, achieving first place. |
| Hepp et al. (2021) | 3D CNN (Adaptation of ResNet) | German National Cohort ($N$ = 10691, ages=20-72) | 3.21 | They used a heteroscedastic noise model to estimate uncertainty, and GradCAM for interpretability. |
| He et al. (2022) | Dual-pathway architecture with transformer network | Validated on CMI and CoRR datasets ($N$ = 8379, ages=0-97) | 2.7 | A global and a local pathway, each with a CNN backbone, are joined with an attention mechanism to fuse global and local information optimally. |
| Leonardsen et al. (2022) | 3D CNN (Variations of SFCN) | T1w MRI scans derived from 21 public datasets ($N$ = 53542, ages=3-95) | 3.9 (in-domain); 5.1 (out-domain) | They used the original SFCN-softmax, as well as a regression variant (with a single output neuron) and a ranking variant. |
| Wood et al. (2022) | 3D CNN (DenseNet 121) | Private multimodal dataset of clinical quality images ($N$ = 23302, ages=18-95) | 3.05 | Guided backpropagation and occlusion sensitivity analysis were performed. |
| Yin et al. (2023) | 3D CNN (one output neuron) | CamCAN, ADNI, HCP, UK Biobank ($N$ = 5851, ages=22-95) | 2.3 for UKB; 4.71 for CamCAN | They provided anatomic maps of brain aging patterns (interpretable). |
| Aqil et al. (2023) | 3D CNN (inspired in Synthmorph) | OpenBHB ($N$ = 3984, ages=6-86) | 4.55 | They integrated diffeomorphic registration with brain age prediction in a unified architecture. |
| Barbano et al. (2023) | 3D CNN (ResNet18, AlexNet and DenseNet121) | OpenBHB ($N$ = 3984, ages=6-86) | **2.61 (in-domain); 3.56 (out-domain)*** | They used contrastive learning techniques to learn domain-invariant features. |
| Gianchandani et al. (2024) | Multi-task U-Net | CamCAN, OASIS and ADNI ($N$ = 651, ages=18-88) | 7.54 | They predicted voxel-level brain age along with two additional tasks: global age prediction and brain tissue segmentation. |

they were able to account for this inherent uncertainty and provide a more comprehensive interpretation of the difference between the predicted age and the actual chronological age.

Additionally, there are some works that employ multi-task learning, in which a single model is trained to address two or more related tasks simultaneously. This strategy takes advantage of the inherent correlations between tasks to potentially improve the performance on each individual task. For instance, the work by Aqil et al. (2023) integrated diffeomorphic registration with brain age prediction within a unified architecture. Similarly, Gianchandani et al. (2024) recently explored a multi-task learning framework with a U-Net that simultaneously predicted voxel-level brain age alongside global age prediction and brain tissue segmentation.

Despite the advancements in deep learning for brain age estimation, most studies still face the challenge of domain shift. It can be noticed that there is significant variability in the results reported by different works depending on the training data. This emphasizes the need for models that generalize well across different domains

and that are able to perform well on unseen data from different acquisition protocols and populations. Some studies address this issue, like the one of Dinsdale et al. (2021b), where they apply the adversarial framework for use in harmonization to obtain domain-invariant feature representations in the context of brain age prediction. However, their application has primarily been limited to a few training and validation domains (three in their case), limiting the practical applicability of their approach. On the other hand, the work by Barbano et al. (2023) introduced contrastive learning techniques for brain age estimation to produce features that were invariant to the domain of acquisition. With this technique, they achieved the state of the art in the OpenBHB dataset (Dufumier et al., 2022), which is highly diverse and multi-site, containing brain scans from 64 different acquisition sites.

Building upon the strengths of existing approaches, this work aims to address some of the challenges in brain age estimation. We employ OpenBHB, a large-scale, multi-site benchmarking dataset to train our models, focusing on achieving generalizability and mitigating domain shift limitations. We also evaluate model performance in sensitive attributes, such as sex, in order to audit and correct potential biases in the trained models. Furthermore, as the training data exhibits a long-tailed distribution with imbalanced age representation, we investigate various techniques to improve model performance on underrepresented age groups. Additionally, we assess the influence of incorporating prior knowledge on the performance of the deep learning models. This involves encoding brain tissue properties using brain tissue segmentation methods and/or local changes by registration to a standard template. Moreover, to provide a comprehensive evaluation framework, we develop machine learning baseline models and compare their performance with the deep learning approaches. Finally, we evaluate the effectiveness of transfer learning strategies to adapt our deep learning models to a separate in-house dataset. By addressing these issues and exploring diverse learning approaches, this work aims to contribute to the advancement of robust and generalizable brain age estimation models.

## 3. Material and methods

### *3.1. Datasets*

#### *3.1.1. OpenBHB*

The Open Big Healthy Brains (OpenBHB) dataset is a publicly available benchmarking dataset specifically designed for brain age prediction with site-effect removal. It aggregates data from ten public sources, including ABIDE 1 and 2, CoRR, GSP, IXI, Localizer, MPI-Leipzig, NAR, NPC, and RBP. OpenBHB focuses exclusively on healthy controls to promote the modelling of normal brain aging (Dufumier et al., 2022).

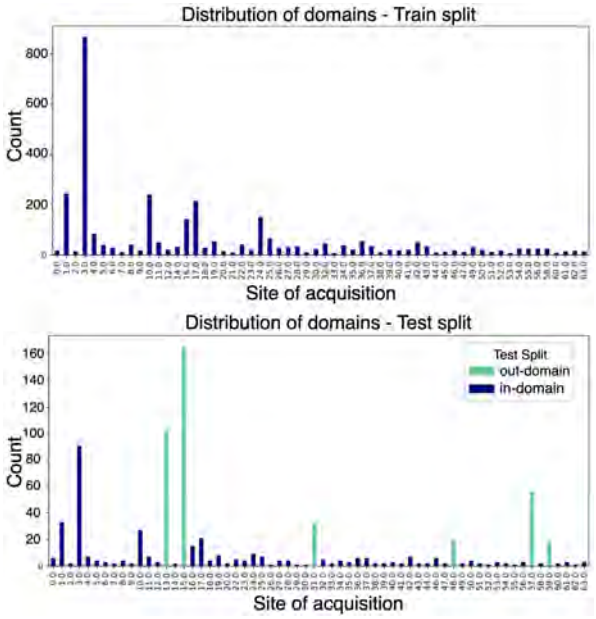



Figure 1: Distribution of domains in the training and test sets of the OpenBHB dataset.

The public part of this dataset comprises $N = 3984$ preprocessed T1w MRI scans, as well as Voxel-Based Morphometry (VBM) volumes (more information on the preprocessing will be detailed in section 3.2). This data originates from 64 different acquisition sites across Europe, America and Asia, promoting generalizability across populations.

The data is provided in two splits: one for training (3227 samples) and one for testing (757 samples). This last split is further divided into in-domain (362 samples coming from the same acquisition sites as training data) and out-domain (395 samples coming from different acquisition sites) subsets (see Figure 1). We further split the provided training set into training (2342 subjects; 72.58%) and validation (885 subjects; 27.42%) sets, while ensuring the same relative distribution of age and acquisition sites across both splits. The age distribution exhibits two main modes centered around 10 and 25 years old, with a long tail extending to 86 years (Figure 2). Sex distribution is well balanced across all age groups (50.17% female and 49.83% male for training and validation; 44.78% female and 55.22% male for testing).

Additionally, associated to this dataset, there is a challenge for brain age estimation. The leaderboard results report the MAE for the in-domain and out-domain subsets of the test set, which allows for better comparison of the developed models. The top performing submission achieves 2.612 years for the MAE of the in-domain set, and 3.564 years for the out-domain set.

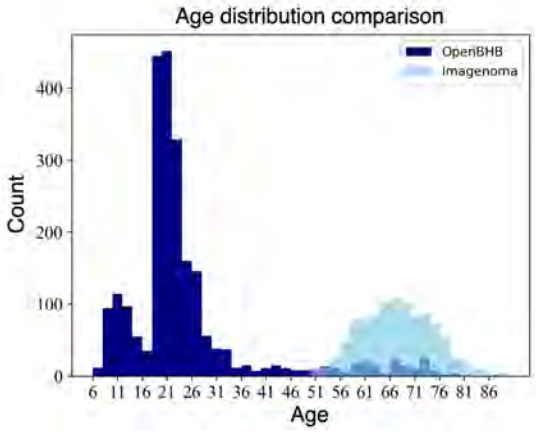



Figure 2: Age distributions of the ImaGenoma and OpenBHB datasets.

#### *3.1.2. ImaGenoma*

The ImaGenoma dataset is an in-house dataset of $N = 1015$ T1w MRI scans, provided by the Hospital Universitari de Girona Doctor Josep Trueta. This dataset offers clinical-grade image quality acquired at 1.5 T field strength. Similar to OpenBHB, only healthy controls were taken into account, meaning that the participating individuals did not show signs of cognitive impairment or any other brain disorders.

The data is divided into training (615 images), validation (202 images), and testing (198 images) sets. Importantly, the splits were constructed to maintain the overall age distribution across all subsets. The age range spans from 50 to 90 years old, with the majority of subjects concentrated between 60 and 70 years old (Figure 2). Sex distribution is relatively balanced, with 46% of female participants and 54% of male participants.

### *3.2. Image Preprocessing*

#### *3.2.1. OpenBHB*

The OpenBHB dataset provides only preprocessed MRI data, and does not offer access to the raw images. The preprocessing pipeline employed for the T1w images of this dataset involved:

- **Bias field correction:** ANTs (Avants et al., 2009) toolbox was used to correct intensity non-uniformities arising from scanner artifacts.
- **Affine registration:** FSL FLIRT (Jenkinson et al., 2002) was employed for affine registration of the T1w images to the Montreal Neurological Institute (MNI) template (Mazziotta et al., 1995) in isotropic 1 mm space. The registration used 9 degrees of freedom, excluding shearing transformations. The MNI template is a standardized brain space derived from averaging a large number of healthy adult MRI scans. This common reference space facilitates the comparison between brain images from different subjects and studies.

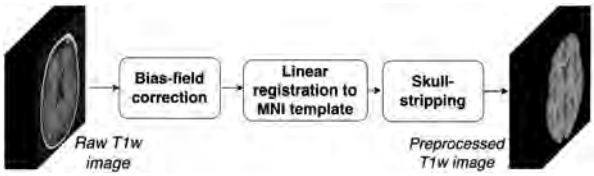



Figure 3: General preprocessing pipeline for T1w images.

Figure 3 shows a flowchart of the main general preprocessing steps, which are common for both ImaGenoma and OpenBHB datasets. An additional preprocessing pipeline was applied to generate VBM volumes using the CAT12 software (Gaser et al., 2022):

- **Nonlinear spatial registration:** A nonlinear transformation was applied to the T1w images to align them to a 1.5 mm isotropic MNI template. Subsequently, the images were resampled to an isotropic resolution of 1 mm to match the dimensions of the other images in the dataset. This ensures consistent voxel sizes across all images for improved analysis.
- **Tissue segmentation:** Gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF) tissues were segmented from the registered images.
- **Intensity normalization:** bias correction of intensity non-uniformities.
- **Modulation:** The segmentation maps were scaled with the amount of volume change introduced by the spatial registration.

It is important to note that only the GM VBM volumes are available; the VBM maps of WM and CSF are not included in the dataset.

#### *3.2.2. ImaGenoma*

A preprocessing pipeline was also applied to the raw T1w images of the ImaGenoma dataset, in order to prepare them for the analysis. The overall pipeline involved:

- **Bias field correction:** The N4 bias field correction method (Tustison et al., 2010) implemented within the SimpleITK library was employed to correct intensity non-uniformities due to scanner artifacts.
- **Linear registration:** FSL's *pairreg* function was utilized for linear registration of the images to a 1 mm isotropic MNI template. This function uses FLIRT (linear) registration (Jenkinson et al., 2002) with 12 degrees of freedom and a special optimization schedule, incorporating two corresponding skull images during the process, which act as a

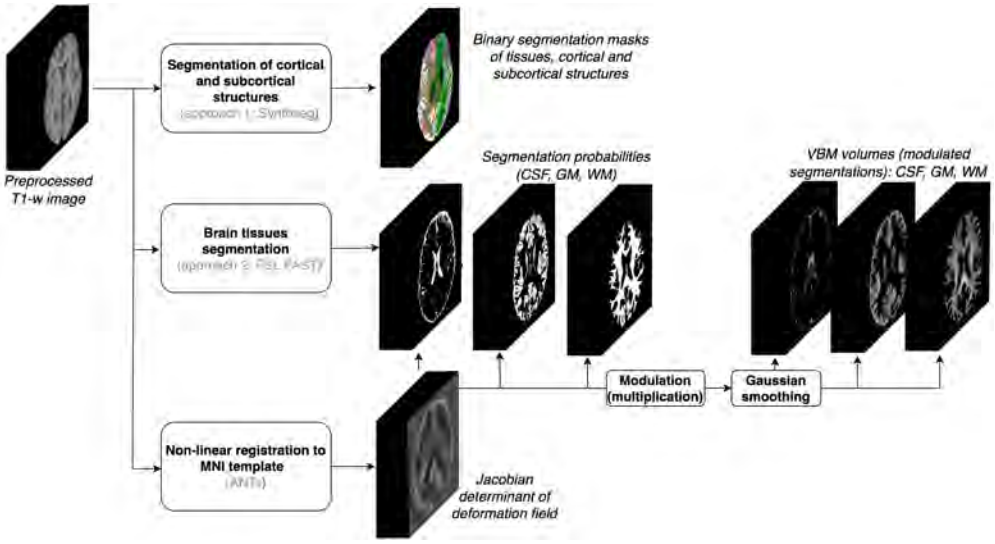



Figure 4: Preprocessing pipeline to obtain custom segmentations and VBM volumes.

reference to maintain consistent scaling throughout the registration.

- **Skull-stripping:** FSL's Brain Extraction Tool (BET) (Smith, 2002) was employed to remove non-brain tissue (skull) from the images.

### *3.2.3. Brain tissue segmentation*

After the initial preprocessing steps, brain tissue segmentation was performed on the images of both OpenBHB and ImaGenoma datasets. These segmentations provided information that was used in deep learning and machine learning approaches, inspired by works such as the ones of Cole et al. (2017), Peng et al. (2021) and More et al. (2023), who included segmentation information for brain age estimation. Firstly, these segmentation masks were employed to extract volumetric features of different cortical and subcortical structures, which were then used to train machine learning models. Secondly, they served as informative priors to guide the training of some of the developed deep learning models. Two different methods were studied to obtain the tissue segmentations:

- **FSL FAST:** This FSL tool was used for probabilistic segmentation of GM, WM and CSF. FAST utilizes a hidden Markov random field model and an Expectation-Maximization algorithm to achieve this task (Zhang et al., 2001). A key advantage of this tool is its ability to provide the probability maps associated to each of the final segmentation labels. These probabilities may offer valuable information regarding the model's confidence in the prediction and thus were later used to train some of the deep learning models.

- **Synthseg:** This convolutional neural network is specifically designed for segmenting brain MRI scans (Billot et al., 2023). It was used because of its robustness across various acquisition protocols, resolutions, subject populations (including healthy young individuals, as well as elderly ones or those with atrophied brains), and variations in preprocessing pipelines. This tool provided binary segmentation masks for 100 different cortical and subcortical structures of the brain. From these segmentations, GM, WM and CSF masks were also obtained.

### *3.2.4. Deformable registration and segmentation modulation*

While the OpenBHB dataset already included VBM volumes, the ImaGenoma dataset did not include them directly. To address this, we opted to generate VBM volumes from the obtained tissue segmentations, followed by modulation of the segmentation probabilities with the information provided by the deformation field. The overall steps of the generation of segmentations and VBM volumes are shown in Figure 4.

The Python library ANTsPy (Avants et al., 2009) was employed for this purpose. For the registration, the Symmetric Normalization transformation was used, which combines affine and deformable transformations, together with a mutual information optimization metric, to achieve the optimal alignment. First, the registration was performed using the T1w MRI as the moving volume, and the 1 mm isotropic MNI template as the fixed volume. Then, the obtained transformation was applied to the segmentations of GM, WM and CSF (obtained with FSL FAST, as explained in section 3.2.3).

Following registration, the Jacobian determinant of the deformable registration's deformation field was obtained, which reflects the local volume changes introduced by the registration process. Then, the registered probabilistic segmentation volumes of GM, WM and

CSF were modulated by multiplying them with the Jacobian determinant. Finally, to improve signal-to-noise ratio, the modulated segmentation probabilities were smoothed using a Gaussian filter with a sigma of 1, as done by Dufumier et al. (2022).

#### *3.2.5. Normalization*

A combined normalization approach was employed specifically for deep learning models. This approach involved dividing each voxel intensity by both the maximum and the mean intensity values within the image (Gong et al., 2021; Peng et al., 2021). This method effectively mitigated the effects of scanner variations and inhomogeneities, ensuring consistency across images, leading to superior performance compared to traditional min-max and z-score normalization methods.

### *3.3. Machine learning approaches*

To establish a baseline for comparison with deep learning models, the performance of various traditional machine learning regression algorithms for brain age estimation was investigated.

The first step consisted in the feature extraction, in which volumetric features were extracted from the brain parcellations generated by the SynthSeg segmentation pipeline. These features consisted of the volumes of 100 distinct brain regions (Billot et al., 2023). Two feature sets were explored. One included the normalized volumes, meaning that each of the 100 regional volumes was divided by the total intracranial volume (TIV). In the other, all 100 regional volumes were used along with the TIV as an additional feature.

Following feature extraction, the data for each dataset (ImaGenoma and OpenBHB) was split into the training and testing sets. Subsequently, standard scaling was applied. This involved fitting the scaler to the training data and then applying the fitted scaler to both the training and testing sets, ensuring consistent scaling across both datasets and avoiding data leakage.

Then, different regression algorithms were evaluated, including:

- **Ridge regression:** This approach utilizes L2 regularization to address overfitting. A grid search was conducted to find the optimal value for the $\alpha$ parameter (options were 0.1, 0.5 and 1.0), which controls the weight of the penalty term on the model's complexity. Higher $\alpha$ values promote simpler models but might reduce their flexibility (Hoerl and Kennard, 1970).
- **Least absolute shrinkage and selection operator (Lasso) regression:** This method implements L1 regularization for feature selection. Similar to ridge regression, a grid search was used to determine the best value for the $\alpha$ parameter. As with ridge regression, a higher $\alpha$ value promotes sparser models with fewer features (Tibshirani, 1996).
- **ElasticNet:** This algorithm combines L1 and L2 regularization, offering advantages of both Lasso and ridge regression. It promotes sparsity and reduces model complexity. A grid search was conducted to identify the optimal hyperparameters for $\alpha$ and the L1 ratio (options were 0.1, 0.5 and 0.9). The $\alpha$ parameter controls the overall amount of regularization, while the L1 ratio balances between L1 and L2 penalties. A value of 1.0 for the L1 ratio corresponds to pure Lasso regression, while a value of 0.0 represents pure ridge regression (Zou and Hastie, 2005).
- **Support Vector Regression:** This kernel-based method was included in the analysis because it offers flexibility in modeling nonlinear relationships. A grid search was employed to tune the hyperparameters C (0.1, 1, 10), $\epsilon$ (0.1, 0.2, 0.5), and kernel type (linear or radial basis function). The C parameter controls the trade-off between maximizing the margin between hyperplanes and minimizing the training error. Higher C values prioritize a large margin but might lead to overfitting. The $\epsilon$ parameter defines the tolerance for misclassified points (a smaller value allows for fewer errors but can increase the model complexity) (Drucker et al., 1996).

Additionally, feature importance was investigated to understand which regional brain volumes were the most influential predictors of brain age.

### *3.4. Deep learning approaches*

In this subsection, we present the various deep learning approaches investigated in our work for brain age estimation from MRI images. Figure 5 summarizes the main methods and techniques explored, including different input types, architectures, and other training strategies. The following subsections will explore the most significant methods and results from these approaches, providing detailed insights into our experimentation and findings.

#### *3.4.1. Simple Fully Convolutional Network (SFCN)*

Our work was mainly based on the use of the SFCN architecture, which was designed by Peng et al. (2021) for brain age estimation. The choice of SFCN as the primary architecture was mainly motivated by two key aspects. Firstly, Peng et al. (2021) provided a pre-trained model trained on a large dataset of over 10,000 scans from the UK Biobank. This pre-trained model serves as a valuable starting point for transfer learning, leveraging the knowledge learned from a vast amount of data to improve performance on potentially smaller datasets. Secondly, the SFCN architecture has been demonstrated by Peng et al. (2021), as well as by others like Gong et al. (2021), to achieve competitive performance in brain age

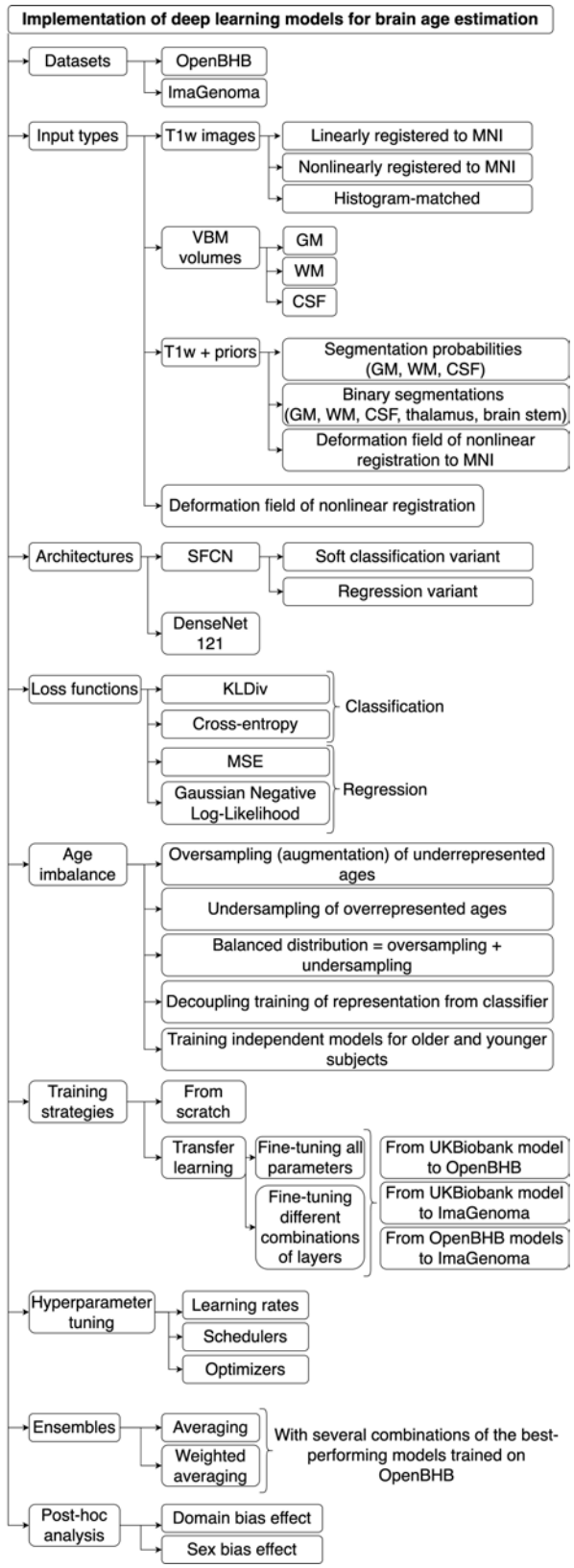



Figure 5: Schematic diagram summarizing the main deep learning approaches investigated in this work.

estimation tasks compared to deeper models. Given its efficiency, pre-trained availability and competitive performance, the SFCN was chosen as the baseline architecture for this study.

Inspired by VGGNet (Simonyan and Zisserman, 2014), the SFCN employs a fully convolutional structure with a significantly reduced number of layers (seven) to minimize computational complexity and memory usage. This design results in a lightweight model containing approximately 3 million parameters, in contrast with deeper architectures like 3D ResNet variants (with tens of millions of parameters) and 2D VGGNet (which has over 100 million parameters).

The SFCN architecture can be conceptually divided into three stages (see Figure 6):

- **Feature extraction (blocks 1-5):** The initial five blocks consist of 3D convolutional layers of $3 \times 3 \times 3$, followed by batch normalization, max pooling ($2 \times 2 \times 2$) and a ReLU activation layer. The input to the network is a 3D image with dimensions $160 \times 192 \times 160$, and the number of channels varies for each case, depending on if priors are added or not. As the image passes through these blocks, feature maps are generated, and the spatial dimensions are progressively reduced, reaching a size of $5 \times 6 \times 5$ after the fifth block. The number of output channels for these blocks are 32, 64, 128, 256, and 256, respectively.

- **Nonlinear mapping (block 6):** The sixth block contains a $1 \times 1 \times 1$ 3D convolutional layer, batch normalization and a ReLU activation layer. This block increases the model's nonlinearity without changing the spatial dimensions of the feature maps, maintaining a size of $5 \times 6 \times 5$ with 64 channels.

- **Age prediction (block 7):** The final block includes an average pooling layer that reduces the spatial dimensions to $1 \times 1 \times 1$, a dropout layer with a 50% dropout rate (used only during training), a fully connected layer, and a softmax output layer.

To predict age, as suggested by Peng et al. (2021), each ground-truth age label was converted into a soft label, a probability distribution centered around the ground truth age. This distribution was modeled as a Gaussian with a user-defined standard deviation ($\sigma$). Consequently, the regression problem was transformed into a multi-class classification problem, where the total age range was divided into bins. For ImaGenoma (age range 50-89), 40 bins of one year each, and $\sigma = 1$, were used. For OpenBHB (age range 6-86), we empirically evaluated two configurations: 40 bins of two years each ($\sigma = 2$), and 80 bins of one year each ($\sigma = 1$). The configuration with 40 bins and $\sigma = 2$ was used for all models, since it achieved better performance. The output of the model was also treated as a probability distribution, with each output node of the model being associated to

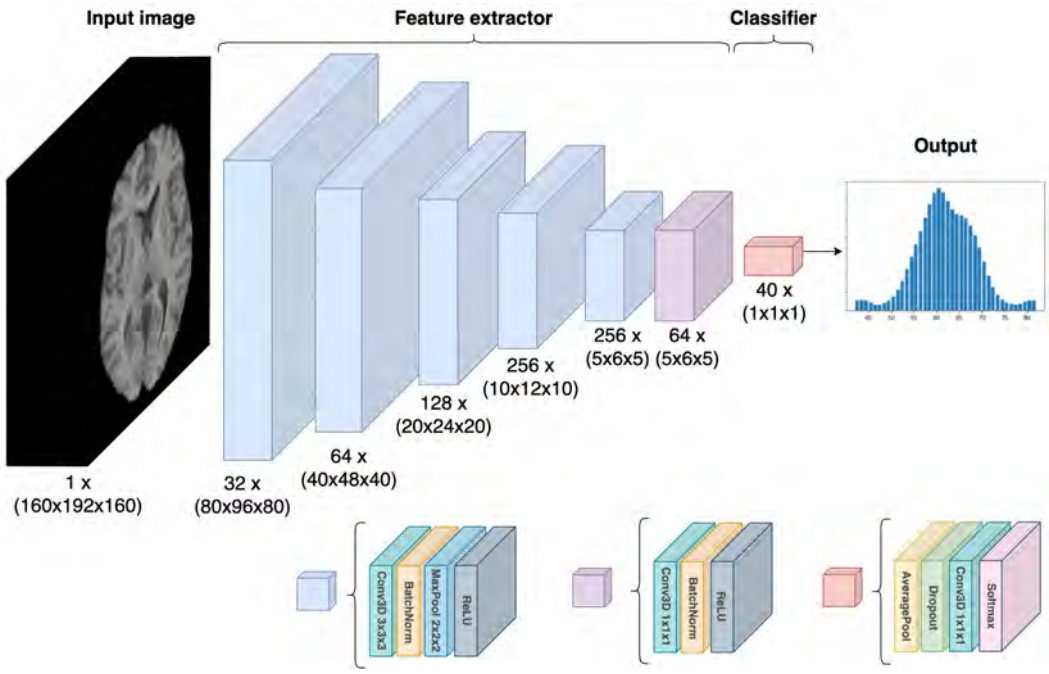



Figure 6: Schematic diagram of the Simple Fully Convolutional Network architecture for brain age estimation.

a bin. The final prediction was obtained from this distribution by calculating the weighted average of each age bin:

$$pred = \sum_{i=1}^{40} x_i \cdot age_i$$

where $x_i$ represents the probability for the $i$-th age interval and $age_i$ is the bin center of the corresponding age bin.

To explore the SFCN architecture's capabilities for direct regression and facilitate the comparison with the original multi-class classification implementation, a modification was made. This involved replacing the final block (block 7) with a single linear layer. The input dimension of this layer matched the number of channels in the last convolutional layer of the feature extractor (64 in this case), while the output dimension was set to 1. This modification enabled the model to directly predict age as a continuous value, rather than a discrete class.

### *3.4.2. Methodological approaches with OpenBHB*

***Baseline model using T1w images***

Several configurations were tested using T1w images linearly registered to the MNI template, as this modality is the standard choice in most existing studies for brain age estimation due to its capability to provide detailed analysis of the anatomical structures and tissues of the brain. These configurations aimed to determine the optimal learning rate, optimizer, and learning rate scheduler, as well as to evaluate different model architectures. In all cases, models were trained from scratch with Xavier initialization of the weights.

The initial configuration employed, as suggested by Peng et al. (2021), the original SFCN with the following configuration: an initial learning rate of $10^{-2}$, a step learning rate scheduler, which decreased the learning rate by a factor of 0.3 every 30 epochs, and a stochastic gradient descent (SGD) optimizer. The loss function used to train the model was the Kullback-Leibler divergence (KLDiv) loss, which was introduced by Kullback (1951) and is a measure of the difference between two probability distributions, given by:

$$KLDiv(P||Q) = \sum_i P(x_i) \cdot \log \frac{P(x_i)}{Q(x_i)}$$

In the context of this work, KLDiv measures the difference between the predicted probability distribution of the subject's age (Q(x)) and the soft label distribution centered around the ground truth age (P(x)). Minimizing this soft-classification loss encourages the model to produce predictions that closely resemble the true age distribution (Peng et al., 2021).

We then explored modifying this approach to identify the optimal hyperparameters. First, the optimizer was changed to Adam. Next, the OneCycle learning rate scheduler was tested, which adjusts the learning rate dynamically, increasing to a peak before decreasing, to potentially improve model convergence and performance. Additionally, the original SFCN was modified to a regression variant, which includes a single output neuron for predicting continuous values and is optimized

using the Mean Square Error (MSE) loss. This variant retained the same initial learning rate and scheduler. Finally, the DenseNet121 architecture, also configured with a single output node and MSE loss, was evaluated using the same initial learning rate and scheduler as the original SFCN setup, since other works like the one of Wood et al. (2022) demonstrated its competitive performance for brain age estimation. The configuration of the original SFCN implementation, with SGD optimizer, KLDiv loss, learning rate of $10^{-2}$ and step learning rate scheduler, was empirically selected as the baseline for the following experiments due to its superior results (as will be seen in the results section).

***Addressing age distribution imbalance***

To overcome the problem of imbalance in the age distribution of the dataset and improve the model's performance across different age groups, three different strategies were employed using the T1w images, linearly registered to the MNI, as input.

The first approach involved oversampling subjects older than 35 years old. Data augmentation techniques were applied for the oversampling, in order to avoid training with identical copies of each image. These included voxel shifting (randomly by 0, 1 or 2 voxels along each axis), randomly flipping around the sagittal plane with a probability of 0.5, and rotating in all three directions by a random angle between 0 and 5 degrees. These augmentations were kept minimal to prevent any structural changes in the image that might have a negative impact on the model's performance.

The second approach combined undersampling of young subjects (younger than 35) with oversampling of older subjects (older than 35) to achieve a completely balanced age distribution for training. The same augmentations used in the previous case were applied to the oversampled subjects.

The third approach was inspired by the approach proposed by Kang et al. (2019), which addresses long-tailed distributed data by decoupling the representation and the classifier during training. This method involved two stages. In the first stage, the model was trained with the imbalanced data. In the second stage, a new model was initialized with the weights from the previously trained model. The feature extractor blocks were frozen, and only the classifier was fine-tuned using balanced data. The balancing was achieved through the same combination of undersampling and oversampling with augmentation as in the second approach.

***Alternative image inputs and priors***

To explore the impact of different image inputs on the performance of brain age estimation models, several models were trained using alternative image representations. These included VBM volumes and various priors added to the T1w images. The use of priors, such as tissue segmentations and probabilistic maps, was investigated to determine if they could help the model focus on relevant anatomical structures and improve overall performance.

The strategies explored are as follows:

1. **GM VBM volume**:
   - The model was trained using the GM VBM volume provided in the OpenBHB dataset as input.
   - The SFCN architecture was used, with training hyperparameters consistent with the baseline model trained with T1w images: an initial learning rate of $10^{-2}$, SGD optimizer, KLDiv loss, and a step learning rate scheduler (decreasing by a factor of 0.3 every 30 epochs).
2. **T1w images with Synthseg binary segmentations**:
   - T1w images linearly registered to the MNI template were used alongside an additional channel containing Synthseg binary segmentations.
   - Separate analyses were conducted for each tissue type: T1w+GM, T1w+WM, and T1w+CSF.
   - Due to the lack of convergence with the KLDiv loss when using multi-channel inputs, the loss function was changed to cross-entropy loss for these analyses.
3. **Custom VBM volumes**:
   - The custom VBM volumes (approach from Figure 4) were used as input to evaluate their impact on model performance.
   - The SFCN architecture, with KLDiv loss and SGD optimizer were used, and the training hyperparameters remained consistent with the ones of the baseline model.
4. **Nonlinearly registered T1w images**:
   - The performance of the model was evaluated using T1w images that were nonlinearly registered to the MNI template.
   - This approach aimed to compare the model's performance relative to training with T1w images that were only linearly registered to MNI.

***Ensemble models***

To investigate whether combining the predictions of multiple models could enhance performance and improve generalization capabilities, several ensemble experiments were conducted. The idea was to average the predictions of different models and evaluate their collective performance.

The models considered for the ensembles were:

- The baseline model trained with the original SFCN and the T1w linearly registered to the MNI template.
- The model trained with the GM VBM.
- The model trained with the regression variant of the SFCN and the T1w linearly registered to the MNI template.
- The model trained with data that had balanced age distribution.
- Models trained with binary segmentation priors (CSF, WM and GM).
- The model trained with the T1w nonlinearly registered to the MNI.
- The model that addressed long-tailed distributed data by decoupling the representation and the classifier during training (with T1w as input).

All possible combinations of these models were evaluated to identify the ensemble that achieved the lowest MAE. We also explored weighted averaging to see if it could further improve the performance of the ensemble.

#### *3.4.3. Methodological approaches with ImaGenoma*

In this section, the methodologies employed on the separate private dataset of clinical-quality images, ImaGenoma, are described. This dataset is smaller than OpenBHB and has a more limited age range, as described in section 3.1.2.

***Training from scratch***

As a baseline, the original SFCN implementation was trained from scratch using the ImaGenoma dataset. The implementation of SFCN with 40 output neurons was used. The initial learning rate was set to $10^{-2}$, with a OneCycle learning rate scheduler, KLDiv loss and Adam optimizer. These hyperparameters were chosen after tuning to find the best configuration.

***Transfer learning from UK-Biobank pretrained model***

To leverage the pretrained model provided by Peng et al. (2021), which was trained on over 10,000 images of the UK Biobank with the same age range as ImaGenoma, several transfer learning approaches were applied. The goal was to determine if the pretrained model could be adapted to be used with the ImaGenoma dataset and to compare the performance of this approach against training the model from scratch with ImaGenoma images. The approaches included fine-tuning only the classifier of the pretrained model, fine-tuning the classifier and the last layer of the feature extractor, and fine-tuning the classifier and the last two layers of the feature extractor.

For these strategies, the Adam optimizer, an initial learning rate of $10^{-2}$, and a step learning rate scheduler (decreasing by a factor of 0.3 every 30 epochs) were used. Since fine-tuning the classifier and the last layer of the feature extractor yielded the best results, an additional approach was tested using the OneCycle learning rate scheduler. This was to see if this scheduler could help the model adapt faster to the new domain and improve the final performance.

***Transfer learning from models trained on OpenBHB***

Further transfer learning techniques were applied using the best performing models trained on the OpenBHB dataset. The aim was to see if models could adapt to the ImaGenoma dataset, despite the age range of ImaGenoma coinciding with the less represented ages in OpenBHB. Several approaches were evaluated, each tested for fine-tuning only the classifier, the classifier plus the last layer of the feature extractor, and the classifier plus the last two layers of the feature extractor. However, only the best-performing combination will be reported in the results section, in order to compare it with the approaches of training from scratch or transfer learning from the model pre-trained on OpenBHB.

By testing these transfer learning approaches, the objective was to determine which method best adapted to the ImaGenoma dataset and whether transferring knowledge from larger datasets, like UK Biobank or OpenBHB, could enhance performance on the images of the smaller in-house dataset.

### *3.5. Implementation details*

Deep learning models were implemented using PyTorch (Version: 2.0.1) (Paszke et al., 2019) and PyTorch Lightning (Version: 2.2.0.post0) (Falcon, William and The PyTorch Lightning team) libraries, within a Python 3.10.12 environment. Scikit-learn (Version: 1.3.0) (Pedregosa et al., 2011) was used for the implementation of traditional machine learning models. Training and evaluation of deep learning models was performed with an NVIDIA A30 GPU with 24GB of memory and CUDA version 12.2.

### *3.6. Evaluation measures*

The performance of the brain age estimation models was evaluated using correlation plots and the following measures:

- **Mean Absolute Error (MAE):** The MAE measures the average absolute difference between the predicted brain age and the actual chronological age of the subjects (Willmott and Matsuura, 2005).
- **Coefficient of determination or R-squared ($R^2$):** $R^2$ represents the proportion of variance in the actual chronological ages that can be explained by the predicted brain ages (James et al., 2013).

Table 3: Performance of traditional machine learning algorithms.

| Dataset | Feature set | Algorithm | MAE (years) | $R^2$ |
|---|---|---|---|---|
| OpenBHB | Normalized | Ridge | 4.93 | 0.80 |
| | | Lasso | 4.91 | 0.80 |
| | | ElasticNet | 4.94 | 0.80 |
| | | SVR | 4.21 | 0.82 |
| | TIV included | Ridge | 5.02 | 0.79 |
| | | Lasso | 5.03 | 0.79 |
| | | ElasticNet | 5.03 | 0.78 |
| | | SVR | **4.18** | **0.82** |
| ImaGenoma | Normalized | Ridge | 4.81 | 0.38 |
| | | Lasso | 4.66 | 0.42 |
| | | ElasticNet | 4.70 | 0.42 |
| | | SVR | 4.94 | 0.33 |
| | TIV included | Ridge | 4.75 | 0.40 |
| | | Lasso | **4.61** | **0.44** |
| | | ElasticNet | 4.62 | 0.44 |
| | | SVR | 4.75 | 0.41 |

- **Pearson's Correlation Coefficient ($r$):** $r$ measures the linear correlation between the predicted and actual brain ages. A value closer to 1 indicates a strong positive linear relationship between the predicted and actual ages, whereas a value of 0 signifies no linear correlation (James et al., 2013).

Additionally, paired or independent samples t-tests were employed to assess statistically significant differences between the prediction errors of our models across different groups (e.g., men vs. women) or approaches.

## 4. Results

### *4.1. Machine learning approaches*

We evaluated the machine learning algorithms on both the ImaGenoma and OpenBHB datasets with two feature sets (dividing each regional volume by the TIV, and using the TIV as extra feature).

Table 3 summarizes the performance metrics (MAE and $R^2$) achieved by the algorithms on both datasets with the two feature sets.

It can be observed that Lasso regression, with the feature set that included the TIV as an extra feature, achieved the best performance with a MAE of 4.61 years and $R^2$ of 0.44 in the ImaGenoma dataset. The optimal $\alpha$ hyperparameter, found with grid search, was 0.1. On the other hand, the method that yielded the best results in the OpenBHB dataset was SVR, with a MAE of 4.18 years and $R^2$ of 0.82 (also including the TIV as an extra feature). The best hyperparameters for this were C = 10, $\epsilon = 0.1$ and the nonlinear kernel.

Analyzing the top features identified through feature importance revealed a consistent pattern across algorithms and datasets. The volumes of the left thalamus, right cerebral cortex and brain stem emerged as the most important predictors of brain age according to ridge regression, Lasso regression and ElasticNet. Additionally, the volume of the third and fourth ventricles consistently held significant importance across all algorithms and datasets.

### *4.2. Deep learning approaches on OpenBHB dataset*

#### *4.2.1. Baseline model with T1w images*

As described in Section 3.4.2, several configurations were evaluated, focusing on hyperparameter tuning, optimizer selection, learning rate scheduling, and architecture choice. The performance of each configuration was assessed using the MAE on the entire test set (in-domain and out-domain combined), the in-domain MAE, the out-domain MAE, the $R^2$ and the correlation coefficient.

Table 4 summarizes the hyperparameters, architectures, and performance metrics of the evaluated models (a numerical ID is assigned in the first column to each experiment for easier future reference). As shown in this table, the configuration employing SGD optimizer, a learning rate of $10^{-2}$, a step learning rate scheduler and the original SFCN architecture with KLDiv loss achieved the best overall performance. Because of this, it was chosen as the baseline for the rest of the experiments.

#### *4.2.2. Addressing age distribution imbalance*

To evaluate the effectiveness of the strategies used to address the imbalance in the age distribution, the models were assessed using the same metrics as the baseline model. Additionally, we analyzed the MAE across different age ranges to understand how well each strategy mitigated the imbalance issue, particularly for older subjects. The results of these evaluations are summarized in Table 5, and the correlation and boxplots for the different strategies are presented in Figure 7.

Table 4: Performance of deep learning models with T1w image inputs.

| ID | Optimizer | LR | Architecture | Loss | MAE (all) | MAE (in) | MAE (out) | $R^2$ | $r$ |
|---|---|---|---|---|---|---|---|---|---|
| 01 | SGD | $10^{-2}$ (step scheduler) | SFCN (classification) | KLDiv | **3.08** | **2.63** | **3.50** | **0.86** | **0.93** |
| 02 | Adam | $10^{-2}$ (step scheduler) | SFCN (classification) | KLDiv | 3.72 | 3.03 | 4.34 | 0.77 | 0.88 |
| 03 | SGD | $10^{-2}$ (OneCycle scheduler) | SFCN (classification) | KLDiv | 3.09 | 2.75 | 3.4 | 0.85 | 0.92 |
| 04 | Adam | $10^{-4}$ (step scheduler) | SFCN (regression) | MSE | 3.28 | 2.95 | 3.57 | 0.84 | 0.92 |
| 05 | Adam | $10^{-4}$ (step scheduler) | DenseNet121 | MSE | 3.37 | 2.87 | 3.83 | 0.83 | 0.91 |

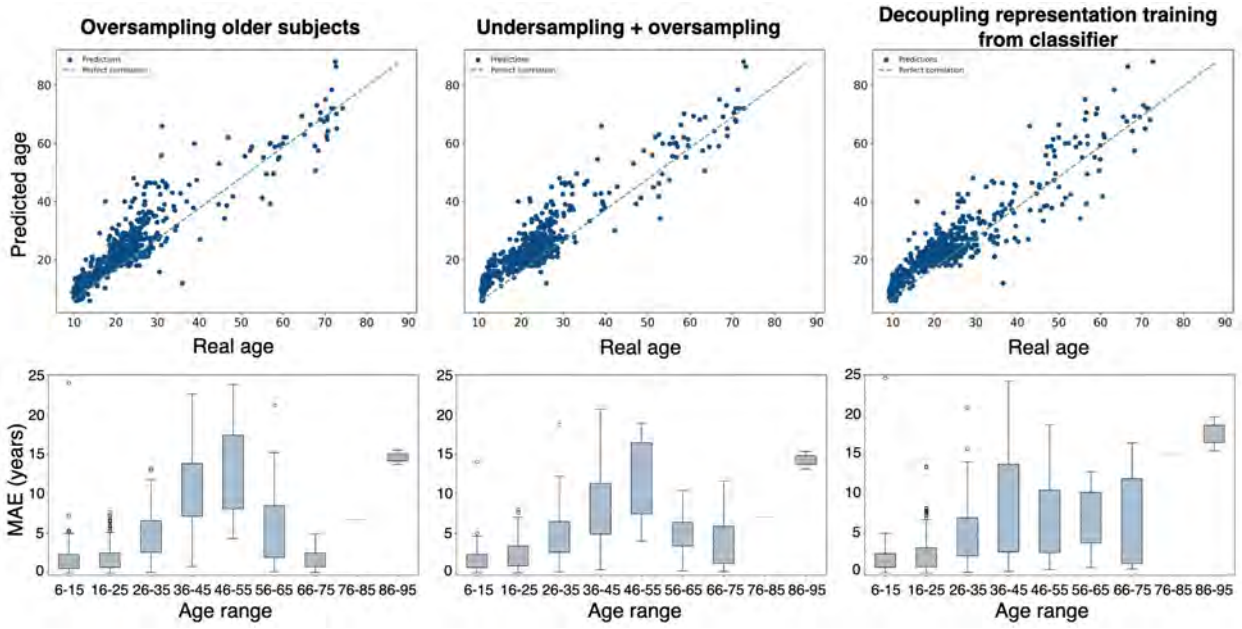



Figure 7: Correlation plot between the actual and predicted ages for all samples of the test set (top) and boxplot with absolute errors per age decade (bottom) for the models trained to mitigate age imbalance.

The results in Table 5 show that oversampling with augmentation (experiment 06) achieved the lowest overall MAE on the entire test set. However, examining the boxplots, it can be observed that while this strategy reduced the error for older subjects, it also resulted in higher error in the middle age range. The same can be observed in the case of undersampling younger subjects and oversampling older ones (experiment 07). On the other hand, the last strategy of decoupling the training of the representation from the classifier, exhibited a more uniform decrease in MAE across most age groups, although the error for the oldest subjects was slightly higher compared to experiments 06 and 07.

*4.2.3. Alternative image inputs and priors*

As mentioned in Section 3.4.2, we investigated the use of VBM volumes and incorporating priors with T1w images as input, as well as using the nonlinearly registered T1w images.

Table 6 summarizes the performance of the models trained with different image inputs. The model using the GM VBM volume provided in the OpenBHB dataset achieved a MAE of 2.97 on the entire test set, demonstrating the best performance in terms of MAE, $R^2$ and correlation coefficient. Among the models trained with T1w images and segmentation priors, the one using CSF segmentation resulted in the lowest overall MAE (3.18), followed by the one using GM (MAE = 3.34). Using custom VBM volumes achieved comparable, but poorer, performance compared to the baseline model. Finally, using T1w images nonlinearly registered to the MNI template resulted in higher error when compared to the model trained with linearly registered T1w images.

The superiority of using only GM VBM volumes for brain age estimation compared to relying on T1w images and priors is further corroborated by the significant difference in performance between experiment 09

Table 5: Summary of results of the different strategies to mitigate age imbalance.

| ID | Strategy | MAE (all) | MAE (in) | MAE (out) | $R^2$ | $r$ |
|---|---|---|---|---|---|---|
| 01 | Baseline model | 3.08 | 2.63 | 3.50 | 0.86 | 0.93 |
| 06 | Oversampling subjects older than 35 | 3.10 | 2.65 | 3.52 | 0.85 | 0.92 |
| 07 | Combined undersampling and oversampling | 3.21 | 2.71 | 3.68 | 0.87 | 0.93 |
| 08 | Decoupling representation and classifier | **3.18** | **2.82** | **3.52** | **0.86** | **0.93** |

Table 6: Performance of deep learning models with alternative image inputs and priors.

| ID | Input | MAE (all) | MAE (in) | MAE (out) | $R^2$ | $r$ |
|---|---|---|---|---|---|---|
| 01 | Baseline model | 3.08 | 2.63 | 3.50 | 0.86 | 0.93 |
| 09 | GM VBM (SPM) | **2.97** | **2.66** | **3.25** | **0.88** | **0.94** |
| 10 | T1w + GM binary segmentation | 3.34 | 2.79 | 3.84 | 0.83 | 0.91 |
| 11 | T1w + WM binary segmentation | 3.62 | 2.90 | 4.28 | 0.76 | 0.88 |
| 12 | T1w + CSF binary segmentation | 3.18 | 2.77 | 3.55 | 0.83 | 0.92 |
| 13 | Custom GM VBM | 3.28 | 2.75 | 3.77 | 0.83 | 0.92 |
| 14 | Custom WM VBM | 3.17 | 2.80 | 3.51 | 0.85 | 0.93 |
| 15 | Custom CSF VBM | 3.25 | 2.85 | 3.61 | 0.84 | 0.92 |
| 16 | T1w nonlinearly registered to MNI | 3.35 | 2.70 | 3.94 | 0.81 | 0.91 |

(trained with GM VBM) and experiment 11 (trained with T1w and WM segmentation). The first model achieved a significantly lower MAE (2.97 years) compared to the latter (3.58 years) (p-value = 0.0001). This statistically significant improvement ($p < 0.001$) highlights the value of GM VBM volumes in capturing relevant information for age prediction.

*4.2.4. Analysis of domain and sex effects*

To assess potential domain biases in the extracted features, we employed t-SNE visualizations (Van der Maaten and Hinton, 2008). These visualizations project high-dimensional features, extracted from the last layer of the feature extractor of the network (purple block of Figure 6), onto a 2D plane, allowing for the exploration of potential clustering patterns based on domain (identified in Figure 8 by color and shape). We selected for this analysis three models trained with different types of inputs, in order to investigate the domain bias introduced by each one. These included:

- Experiment 08 (T1w input): We aimed to determine if the intensity variations across scanners, present in T1w images, introduced significant domain bias in the extracted features.
- Experiment 12 (T1w + CSF binary segmentation prior): We investigated whether including segmentation prior information influences the domain-invariance of the features compared to using only T1w images. This specific model was chosen for the analysis as it demonstrated the best performance among the models trained with segmentation priors.
- Experiment 09 (GM VBM input):We compared this model to the T1w-based models to assess if employing VBM volumes reduces domain bias in the learned features.

When examining the t-SNE plots of Figure 8 for models trained on T1w images (both with and without priors), data points from out-of-domain sites (like sites 13, 15, or 57) are easily identifiable and clustered, suggesting a potential bias. In contrast, the t-SNE plot for the VBM-based model reveals more diffuse clusters, with features covering a wider portion of the space.

The influence of sex on the model's performance was evaluated using a two-sided t-test comparing the absolute errors between male and female subjects for the best model (experiment 09). Moreover, similar to the domain analysis, a t-SNE visualization plot was created, but colored by sex (male/female) instead of domain. This visualization aimed to identify any potential clustering patterns based on sex within the feature space. For the model trained with GM VBM, the t-test resulted in a p-value of 0.22, suggesting no statistically significant difference (at a significance level of 0.05) in MAE between males and females. Also when visualizing the tSNE plot of Figure 9, no clusters can be identified based on sex.

*4.2.5. Ensembles*

To assess the potential of ensemble learning for brain age estimation, we combined predictions from various models explored in Section 3.4.2. After evaluating different combinations, an ensemble achieved the best performance using the following models: the model trained with decoupled representation and classifier (exp 08),

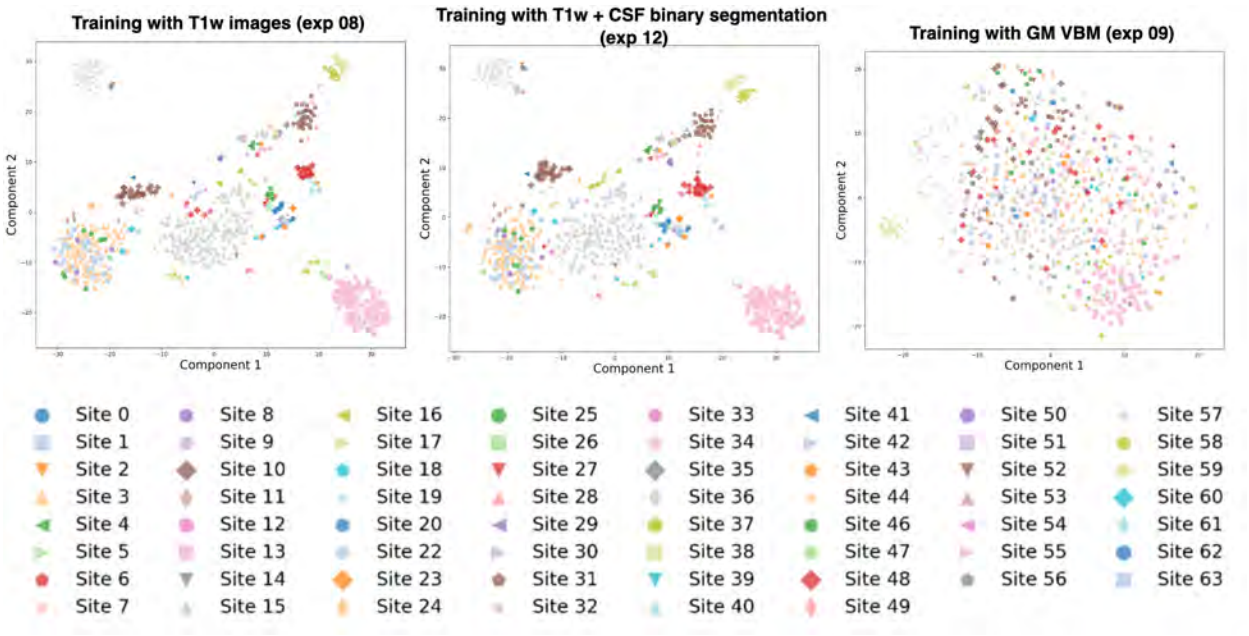



Figure 8: t-SNE visualization of features for experiments 08, 12, and 09, colored and shaped by domain.

Table 7: Performance of ensemble models in comparison with the best-performing individual deep learning model.

| ID | Strategy | MAE (all) | MAE (in) | MAE (out) | $R^2$ | $r$ |
|---|---|---|---|---|---|---|
| 09 | GM VBM (OpenBHB) | 2.97 | 2.66 | 3.25 | **0.88** | 0.94 |
| 17 | Non-weighted average ensemble | 2.74 | 2.47 | 3.00 | 0.87 | 0.95 |
| 18 | Weighted average ensemble | **2.70** | **2.39** | **2.99** | 0.87 | **0.95** |

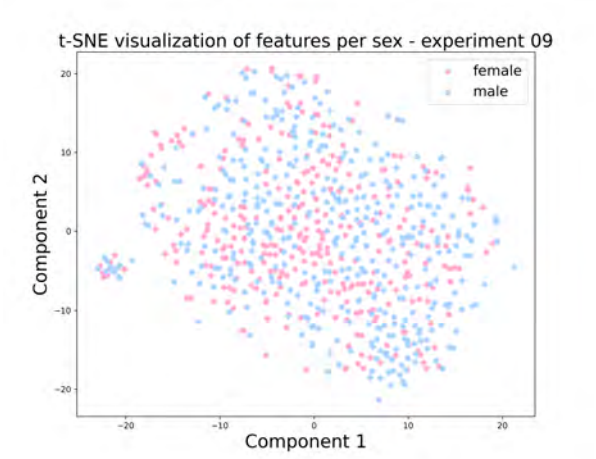



Figure 9: t-SNE visualization of features for experiment 09, colored by sex.

the model trained with GM VBM (exp 09), the model trained with T1w and oversampling/augmentation of older subjects (exp 06), and the model trained with SFCN regression variant (exp 04). The initial ensemble employed a non-weighted average of the predictions of these individual models, achieving a general MAE of 2.74 years.

Further optimization was achieved through a weighted ensemble approach. Weights for each model were identified using a function that explored all possible normalized combinations between 0 and 1. The resulting optimal weights assigned the highest importance to the GM VBM model (experiment 09, weight: 0.354), followed by the model with oversampling and T1w input (experiment 06, weight: 0.27), and equal weights (0.188 each) for the models of experiments 08 (trained with decoupled representation and classifier) and 04 (using SFCN regression variant).

Table 7 shows the results of the best individual model (experiment 09, trained with GM VBM), compared to the performance of the non-weighted and weighted average ensembles. It can be observed that the weighted ensemble slightly outperformed the non-weighted version, and that both ensembles outperformed the individual models. The weighted ensemble achieved a general MAE of 2.70 years, which was a statistically significant improvement (p-value=0.003) compared to the best individual model trained with GM VBM volumes. This demonstrates the effectiveness of ensemble learning in enhancing brain age estimation performance.

Figure 10 visually shows the performance of the final weighted ensemble through correlation plots between the predicted and actual ages for all the test set, and the in-domain and out-domain subsets.

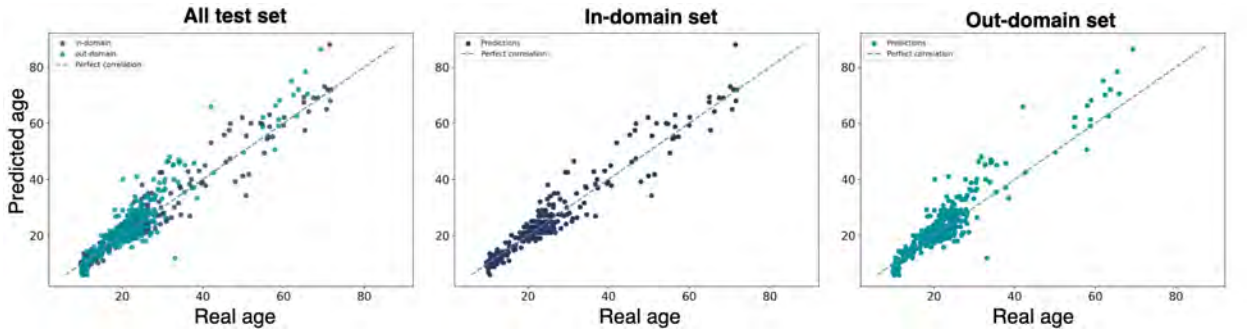



Figure 10: Correlation plot between the actual and predicted ages for the entire test set (left), the in-domain subset (middle), and the out-domain subset (right) of the ensemble model.

### *4.3. Deep learning approaches on ImaGenoma dataset*

The performance of some SFCN models, trained with different strategies on the ImaGenoma dataset as explained on Section 3.4.2, was evaluated. Three main training strategies were assessed: training from scratch, transfer learning from the pre-trained model on UK Biobank dataset, and transfer learning from the best performing models trained on the OpenBHB dataset.

For the transfer learning experiments, in the case of fine-tuning the model pre-trained on UK Biobank, the best-performing model (MAE of 3.28 years) was obtained by fine-tuning only the last layer of the feature extractor and the classifier block. When applying transfer learning from the model of experiment 01 trained on OpenBHB, the last two layers of the feature extractor and the classifier needed to be fine-tuned, achieving a MAE of 3.66 years.

Table 8 summarizes the performance metrics (MAE, $R^2$ and correlation coefficient) for each strategy. From this table, it can be seen that the best performing model was obtained by fine-tuning the model that was pre-trained on the UK Biobank dataset. The correlation plot between the real ages and the ages predicted by this model can be seen in Figure 11.

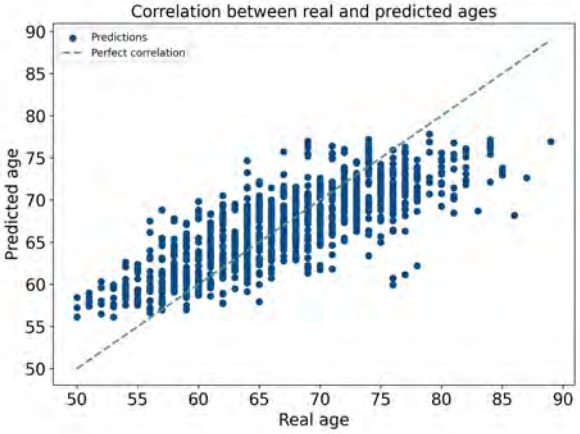



Figure 11: Correlation plot between the actual and predicted ages with the best-performing model (transfer learning from model pre-trained on UK Biobank) for the ImaGenoma test set.

## 5. Discussion

This study aimed to develop and evaluate different methods for accurate and robust brain age prediction, leveraging a large-scale, multi-site dataset and exploring both traditional machine learning and deep learning approaches. We also investigated the impact of incorporating prior knowledge and the effectiveness of transfer learning techniques. Our findings provide insights into the relative strengths and limitations of these approaches and their potential for improving brain age estimation.

Our evaluation of traditional machine learning regression algorithms for brain age estimation yielded competitive results compared to recent works in the field. Specifically, the Lasso regression model achieved a MAE of 4.61 years on the ImaGenoma dataset, while SVR obtained a MAE of 4.18 years on the OpenBHB dataset. De Lange et al. (2022) had reported a MAE of 4.18 years on the UKBiobank, which has a similar age range to ImaGenoma but with a larger sample size, while Da Costa et al. (2020) had achieved a MAE of 4.571 years on the PAC2019 dataset, which is comparable to OpenBHB in terms of age distribution and sample size. This indicated that volumetric measures from cortical and subcortical structures capture valuable information related to brain aging, particulary in structures like the left thalamus, right cerebral cortex, brainstem, and ventricles. These findings align with previous studies by Fama and Sullivan (2015) who reported thalamic and cortical gray matter volume decline with age, Luft (1999) who observed brainstem volume decrease after the age of 50, and Apostolova et al. (2012) who linked ventricular enlargement to aging. Interestingly, a linear model (Lasso) performed better on the ImaGenoma dataset, while a nonlinear model (SVR) was better on the OpenBHB dataset. This suggests that within a limited age span, brain changes might be more linear, effectively captured by Lasso's ability to identify these linear relationships and reduce model complexity. Conversely,

Table 8: Performance of models on ImaGenoma dataset.

| Strategy | Source dataset | MAE | $R^2$ | r |
|---|---|---|---|---|
| Training from scratch | - | 5.15 | 0.21 | 0.27 |
| Transfer learning | UK Biobank | **3.28** | **0.64** | **0.83** |
| | OpenBHB | 3.66 | 0.61 | 0.79 |

for a broader age range with potentially more complex brain changes, nonlinear models like SVR with a nonlinear kernel may be more suitable.

Deep learning models outperformed traditional machine learning approaches, likely due to their ability to directly analyze images and capture nonlinear and complex relationships between voxels that volumetric measures might miss. When establishing baseline deep learning models, our findings suggest that employing a soft classification approach with SFCN networks yield better results compared to direct regression. This could be attributed to the tendency of regression models to predict values closer to the training data's average, potentially affecting the performance negatively, especially with imbalanced datasets. The soft-classification approach might mitigate this issue and capture more nonlinear relationships.

We investigated various strategies to address the imbalanced age distribution within the datasets. All three methods improved performance for under-represented older subjects without significantly compromising accuracy for younger ones. Oversampling with augmentation of older subjects achieved the lowest overall MAE, but boxplots revealed that error for middle-aged subjects remained high. Decoupling representation training from the classifier achieved a more uniform decrease in MAE across most age groups. This suggests that oversampling might lead to overfitting to the characteristics of the oversampled data, hindering generalizability. Decoupling the training stages and using balanced data for fine-tuning the classifier might allow the model to learn more generalizable representations applicable to a wider age range.

The most successful deep learning models utilized GM VBM volumes as input. This indicates that the GM segmentation, and the deformation field information with which it is modulated, provide crucial details that aid the model in differentiating subjects of different ages. This aligns with the established knowledge of age-related GM atrophy (Oh et al., 2014). Additionally, training with VBM volumes reduced the error in the out-domain test set subset, as these volumes are less susceptible to scanner or protocol variations compared to T1w images. In contrast, incorporating segmentation priors as input did not significantly improve performance, suggesting the model primarily relied on T1w image information for age prediction. Training with custom VBM volumes yielded lower performance compared to those provided by the OpenBHB dataset. This could be due to segmentation quality or the fact that our nonlinear registration to the MNI template started from a pre-registered T1w image, potentially missing crucial deformation details present in the original image. Furthermore, training with nonlinearly registered T1w images resulted in worse performance compared to linear registration. This is reasonable since linear registration introduces less deformation, potentially preserving tissue and structural details relevant for age prediction.

Ensembling predictions from various models yielded superior performance compared to individual models. The weighted ensemble, which accounted for the strengths and weaknesses of individual models, achieved the best performance with a MAE of 2.70 years. This approach improved generalizability, reflected in the reduced gap between in-domain and out-domain MAE. Specifically, our best individual model achieved MAE of 2.66 in the in-domain subset and 3.25 in the out-domain subset, while the ensemble further improved these results to 2.39 and 2.99, respectively. To the best of our knowledge, these results show better performance than the state-of-the-art method in the OpenBHB dataset proposed by Barbano et al. (2023), who achieved MAEs of 2.61 and 3.56 in the same subsets.

Finally, regarding the models for the ImaGenoma dataset, training a deep learning model from scratch on this data resulted in high error (MAE of 5.15 years). This can be attributed to the limited dataset size and age range, and potentially lower image quality. Transfer learning from models trained on larger datasets significantly improved performance on this clinical dataset. The model transferred from OpenBHB achieved an error of 3.66 years despite the limited overlap in the age between both datasets. The best performance was achieved by fine-tuning the model pre-trained on UK Biobank (MAE of 3.28 years), which had a substantial number of images in the same age range as ImaGenoma, highlighting the effectiveness of leveraging large datasets for improving performance on smaller datasets.

One of the limitations of our work is related to the data used for training. Even though the OpenBHB dataset, which was the primary dataset for developing our models, is a large and diverse dataset, it is highly imbalanced in terms of age distribution. Consequently, our models do not perform as well for older subjects due to the limited number of images in this age range. This is a significant limitation, as many neurodegen-

erative diseases, which we aim to predict using estimated brain age, predominantly affect older individuals. In future work, we plan to incorporate more data from middle-aged and older subjects to develop models that perform well across all age groups. Moreover, another limitation is the lack of interpretability. To address this limitation, future work will explore techniques for explainable deep learning, in order to understand which features in the brain scans contribute most significantly to the age prediction, and assess whether these regions align with findings from our machine learning analyses. Additionally, we aim to evaluate and improve brain age estimation models for diseased subjects that could present brain lesions (e.g., subjects with multiple sclerosis), validating whether brain age delta could serve as a biomarker for cognitive decline and neurodegeneration.

## 6. Conclusions

This study demonstrates the potential of both traditional machine learning and deep learning approaches for brain age estimation. We implemented and evaluated a comprehensive range of methods for brain age estimation, exploring the influence of different input types (segmentations, VBM volumes and priors) and data balancing strategies. Additionally, we investigated transfer learning to enhance performance on a smaller in-house dataset. Traditional machine learning models, while effective, were outperformed by deep learning models, particularly those leveraging VBM volumes and ensemble methods. Brain age estimation using GM VBM volumes achieved the best performance. Ensemble learning further improved this performance, demonstrating the value of combining different models' strengths. Importantly, our method achieved superior performance on the OpenBHB benchmark dataset compared to the previously reported state-of-the-art method. These findings contribute to the advancement of brain age estimation models, offering valuable insights for future research and clinical applications.

## Acknowledgments

I would like to express my gratitude to Hospital Universitari Dr. Josep Trueta for providing the images of the ImaGenoma dataset. Special thanks to my supervisors, Adrià Casamitjana, Arnau Oliver and Xavier Lladó, for their invaluable help, guidance and patience throughout this project. I am also deeply grateful to my family and friends for their constant support during my MAIA journey.

## References

Apostolova, L.G., Green, A.E., Babakchanian, S., Hwang, K.S., Chou, Y.Y., Toga, A.W., Thompson, P.M., 2012. Hippocampal atrophy and ventricular enlargement in normal aging, mild cognitive impairment (mci), and alzheimer disease. Alzheimer Disease & Associated Disorders 26, 17–27.

Aqil, K.H., Kulkarni, T., Jayakumar, J., Ram, K., Sivaprakasam, M., 2023. Confounding factors mitigation in brain age prediction using mri with deformation fields, in: Predictive Intelligence in Medicine, Springer Nature Switzerland, Cham. p. 58–69.

Avants, B.B., Tustison, N., Song, G., et al., 2009. Advanced normalization tools (ants). Insight j 2, 1–35.

Barbano, C.A., Dufumier, B., Duchesnay, E., Grangetto, M., Gori, P., 2023. Contrastive learning for regression in multi-site brain age prediction, in: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), p. 1–4.

Bellantuono, L., Marzano, L., La Rocca, M., Duncan, D., Lombardi, A., Maggipinto, T., Monaco, A., Tangaro, S., Amoroso, N., Bellotti, R., 2021. Predicting brain age with complex networks: From adolescence to adulthood. NeuroImage 225, 117458.

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.

Cole, J.H., Poudel, R.P.K., Tsagkrasoulis, D., Caan, M.W.A., Steves, C., Spector, T.D., Montana, G., 2017. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. NeuroImage 163, 115–124.

Cole, J.H., Ritchie, S.J., Bastin, M.E., Valdés Hernández, M.C., Muñoz Maniega, S., Royle, N., Corley, J., Pattie, A., Harris, S.E., Zhang, Q., Wray, N.R., Redmond, P., Marioni, R.E., Starr, J.M., Cox, S.R., Wardlaw, J.M., Sharp, D.J., Deary, I.J., 2018. Brain age predicts mortality. Molecular Psychiatry 23, 1385–1392.

Da Costa, P.F., Dafflon, J., Pinaya, W.H.L., 2020. Brain-age prediction using shallow machine learning: Predictive analytics competition 2019. Frontiers in Psychiatry 11, 604478.

De Lange, A.G., Anatürk, M., Rokicki, J., Han, L.K.M., Franke, K., Alnæs, D., Ebmeier, K.P., Draganski, B., Kaufmann, T., Westlye, L.T., Hahn, T., Cole, J.H., 2022. Mind the gap: Performance metric evaluation in brain-age prediction. Human Brain Mapping 43, 3113–3129.

Dinsdale, N.K., Bluemke, E., Smith, S.M., Arya, Z., Vidaurre, D., Jenkinson, M., Namburete, A.I.L., 2021a. Learning patterns of the ageing brain in mri using deep convolutional networks. NeuroImage 224, 117401.

Dinsdale, N.K., Jenkinson, M., Namburete, A.I.L., 2021b. Deep learning-based unlearning of dataset bias for mri harmonisation and confound removal. NeuroImage 228, 117689.

Driscoll, I., Davatzikos, C., An, Y., Wu, X., Shen, D., Kraut, M., Resnick, S.M., 2009. Longitudinal pattern of regional brain volume change differentiates normal aging from mci. Neurology 72, 1906–1913.

Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A., Vapnik, V., 1996. Support vector regression machines, in: Proceedings of the 9th International Conference on Neural Information Processing Systems, MIT Press, Cambridge, MA, USA. p. 155–161.

Dufumier, B., Grigis, A., Victor, J., Ambroise, C., Frouin, V., Duchesnay, E., 2022. Openbhb: a large-scale multi-site brain mri data-set for age prediction and debiasing. NeuroImage 263, 119637.

Falcon, William and The PyTorch Lightning team, . Pytorch lightning (version 1.4). URL: `https://www.pytorchlightning.ai`.

Fama, R., Sullivan, E.V., 2015. Thalamic structures and associated cognitive functions: Relations with age and aging. Neuroscience & Biobehavioral Reviews 54, 29–37.

Franke, K., Ziegler, G., Klöppel, S., Gaser, C., 2010. Estimating the age of healthy subjects from t1-weighted mri scans using kernel methods: Exploring the influence of various parameters. NeuroImage 50, 883–892.

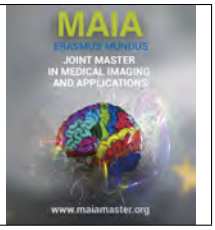
Gaser, C., Dahnke, R., Thompson, P.M., Kurth, F., Luders, E., Initiative, A.D.N., 2022. Cat – a computational anatomy toolbox for the analysis of structural mri data. bioRxiv .

Gaser, C., Franke, K., Klöppel, S., Koutsouleris, N., Sauer, H., Initiative, A.D.N., 2013. Brainage in mild cognitive impaired patients: Predicting the conversion to alzheimer's disease. PLoS ONE 8,

e67346.
Gianchandani, N., Dibaji, M., Ospel, J., Vega, F., Bento, M., MacDonald, M.E., Souza, R., 2024. A voxel-level approach to brain age prediction: A method to assess regional brain aging. Machine Learning for Biomedical Imaging 2, 761–795.
Gong, W., Beckmann, C.F., Vedaldi, A., Smith, S.M., Peng, H., 2021. Optimising a simple fully convolutional network for accurate brain age prediction in the pac 2019 challenge. Frontiers in Psychiatry 12.
He, S., Grant, P.E., Ou, Y., 2022. Global-local transformer for brain age estimation. IEEE Transactions on Medical Imaging 41, 213–224.
Hepp, T., Blum, D., Armanious, K., Schölkopf, B., Stern, D., Yang, B., Gatidis, S., 2021. Uncertainty estimation and explainability in deep learning-based age estimation of the human brain: Results from the german national cohort mri study. Computerized Medical Imaging and Graphics 92, 101967.
Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning: with Applications in R. Springer.
Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17, 825–841.
Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y., 2019. Decoupling representation and classifier for long-tailed recognition. arXiv preprint arXiv:1910.09217 .
Khundrakpam, B.S., Tohka, J., Evans, A.C., 2015. Prediction of brain maturity based on cortical thickness at different spatial resolutions. NeuroImage 111, 350–359.
Koutsouleris, N., Davatzikos, C., Borgwardt, S., Gaser, C., Bottlender, R., Frodl, T., Falkai, P., Riecher-Rossler, A., Moller, H.J., Reiser, M., Pantelis, C., Meisenzahl, E., 2014. Accelerated brain aging in schizophrenia and beyond: A neuroanatomical marker of psychiatric disorders. Schizophrenia Bulletin 40, 1140–1153.
Kullback, S., 1951. Kullback-leibler divergence.
Leonardsen, E.H., Peng, H., Kaufmann, T., Agartz, I., Andreassen, O.A., Celius, E.G., Espeseth, T., Harbo, H.F., Hogestol, E.A., Lange, A.M.d., Marquand, A.F., Vidal-Piñeiro, D., Roe, J.M., Selbaek, G., Sorensen, O., Smith, S.M., Westlye, L.T., Wolfers, T., Wang, Y., 2022. Deep neural networks learn general and clinically relevant representations of the ageing brain. NeuroImage 256, 119210.
Levakov, G., Rosenthal, G., Shelef, I., Raviv, T.R., Avidan, G., 2020. From a deep learning model back to the brain-identifying regional predictors and their relation to aging. Human Brain Mapping 41, 3235–3252.
Luft, A.R., 1999. Patterns of age-related shrinkage in cerebellum and brainstem observed in vivo using three-dimensional mri volumetry. Cerebral Cortex 9, 712–721.
Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-sne. Journal of machine learning research 9.
Madden, D.J., Bennett, I.J., Song, A.W., 2009. Cerebral white matter integrity and cognitive aging: Contributions from diffusion tensor imaging. Neuropsychology Review 19, 415–435.
Mazziotta, J.C., Toga, A.W., Evans, A., Fox, P., Lancaster, J., 1995. A probabilistic atlas of the human brain: Theory and rationale for its development: The international consortium for brain mapping (icbm). Neuroimage 2, 89–101.
More, S., Antonopoulos, G., Hoffstaedter, F., Caspers, J., Eickhoff, S.B., Patil, K.R., 2023. Brain-age prediction: A systematic comparison of machine learning workflows. NeuroImage 270, 119947.
Mwangi, B., Hasan, K.M., Soares, J.C., 2013. Prediction of individual subject's age across the human lifespan using diffusion tensor imaging: A machine learning approach. NeuroImage 75, 58–67.
Nenadić, I., Dietzek, M., Langbein, K., Sauer, H., Gaser, C., 2017. Brainage score indicates accelerated brain aging in schizophrenia, but not bipolar disorder. Psychiatry Research: Neuroimaging 266, 86–89.
Niu, X., Zhang, F., Kounios, J., Liang, H., 2020. Improved prediction of brain age using multimodal neuroimaging data. Human Brain Mapping 41, 1626–1643.
Oh, H., Madison, C., Villeneuve, S., Markley, C., Jagust, W.J., 2014. Association of gray matter atrophy with age, beta-amyloid, and cognition in aging. Cerebral Cortex 24, 1609–1618.
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32. Curran Associates, Inc., pp. 8024–8035.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.
Peng, H., Gong, W., Beckmann, C.F., Vedaldi, A., Smith, S.M., 2021. Accurate brain age prediction with lightweight deep neural networks. Medical Image Analysis 68, 101871.
Salih, A., Boscolo Galazzo, I., Raisi-Estabragh, Z., Rauseo, E., Gkontra, P., Petersen, S.E., Lekadir, K., Altmann, A., Radeva, P., Menegaz, G., 2021. Brain age estimation at tract group level and its association with daily life measures, cardiac risk factors and genetic variants. Scientific Reports 11, 20563.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv1409.1556 .
Smith, S.M., 2002. Fast robust automated brain extraction. Human Brain Mapping 17, 143–155.
Smith, S.M., Vidaurre, D., Alfaro-Almagro, F., Nichols, T.E., Miller, K.L., 2019. Estimation of brain age delta from brain imaging. NeuroImage 200, 528–539.
Soumya Kumari, L.K., Sundarrajan, R., 2024. A review on brain age prediction models. Brain Research 1823, 148668.
Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., Collins, R., 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.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58, 267–288.
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.
Willmott, C., Matsuura, K., 2005. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. Climate Research 30, 79–82.
Wood, D.A., Kafiabadi, S., Busaidi, A.A., Guilhem, E., Montvila, A., Lynch, J., Townend, M., Agarwal, S., Mazumder, A., Barker, G.J., Ourselin, S., Cole, J.H., Booth, T.C., 2022. Accurate brain-age models for routine clinical mri examinations. NeuroImage 249, 118871.
Yin, C., Imms, P., Cheng, M., Amgalan, A., Chowdhury, N.F., Massett, R.J., Chaudhari, N.N., Chen, X., Thompson, P.M., Bogdan, P., Irimia, A., the Alzheimer's Disease Neuroimaging Initiative, 2023. Anatomically interpretable deep learning of brain age captures domain-specific cognitive impairment. Proceedings of the National Academy of Sciences 120, e2214634120.
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.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B: Statistical Methodology 67, 301–320.

Medical Imaging and Applications

Master Thesis, June 2024




# Multifaceted Image Analysis for Cellular Morphology, Neuron Networks, and Protein Expression Segmentation in Bioelectronic Interfaces

Esther Ivanova Matamoros Alcivar[a], Claudia Latte Bovio[b,d,e], Ziyu Gao[b], Francesca Santoro[b,c,d]

[a] *Universitat de Girona, Girona, Spain.*
[b] *Faculty of Electrical Engineering and IT, RWTH Aachen, 52074, Germany.*
[c] *Institute for Biological Information Processing-Bioelectronics, Forschungszentrum Juelich, 52428, Germany.*
[d] *Tissue Electronics, Istituto Italiano di Tecnologia, 80125 Naples, Italy.*
[e] *Dipartimento di Chimica, Materiali e Produzione Industriale, Università di Napoli Federico II, 80125, Naples, Italy*

**Abstract**

In neuroelectronics, the quest for enhanced imaging and analysis techniques is critical, not only for advancing our understanding of neuronal behavior but also for improving bioelectronic interfaces. These interfaces, crucial for capturing the intricate dynamics of neuronal signals, rely on precise imaging to tackle the challenges posed by the complex and unpredictable morphology of neurons, influenced by their inherent polarity.

This study focuses on developing innovative methodologies to refine neuroelectronic tools, aligning to better characterize neuron morphology and interaction. The aim is to translate these advancements into more reliable and effective interfaces that can adapt to the dynamic nature of neuronal structures.

Key achievements include the enhanced image analysis techniques that have allowed for detailed characterization of neuron structures. This has led to the segmentation of clusters of nuclei and neurites, improving our ability to study cellular responses to environmental changes. The measurement of fluorescence intensity across imaging channels has provided insights into neuronal function and protein dynamics related to cell adhesion. By using segmentation and graph analysis, it has improved our understanding of neuronal network dynamics.

These contributions enrich the toolkit for neuroelectronic applications, aiding in the diagnosis and treatment of neurological disorders. This exploratory research highlights the commitment to enhancing neuroelectronic methodologies and sets the stage for future advancements in the field.

*Keywords:* Neuroelectronics, Interfaces, Image analysis, Neuronal behavior, Neuron morphology

## 1. Introduction

Bioelectronics is a multidisciplinary field that merges biology, electronics, and materials science to create devices that interface with biological systems (Cho et al., 2021). These devices, such as biosensors (Martínez et al., 2018) and neural implants (Aspiotis et al., 2022), are designed to interact with biological components like cells and tissues, allowing for the recording, stimulation, and modulation of biological signals (Cuttaz et al., 2024). Bioelectronic devices play a vital role in medical technology, enabling the diagnosis and monitoring of health conditions by detecting various physical, electrophysiological (Kumar et al., 2024), and biochemical markers (Kireev et al., 2022). The field utilizes biocompatible materials to ensure safe integration with biological systems and employs electrical signals to control biological processes, offering transformative applications in healthcare, neuroscience, and beyond (Cuttaz et al., 2024).

### *1.1. Neuroelectronics*

Within the broader scope of bioelectronics, neuroelectronics specifically combines the expertise from neuroscience, electronics, and materials science to innovate technologies that interface with the nervous system (Qi et al., 2023). This interdisciplinary field aims to develop neural interfaces that precisely record and stimulate neural activity, offering significant potential for

treating neurological disorders (Krook-Magnuson et al., 2015).

One of the clinical implications of neuroelectronics is its capacity for therapeutic neuromodulation. Devices such as those used in deep brain stimulation can effectively manage symptoms of Parkinson's disease (Neumann et al., 2023), epilepsy (Ganguli et al., 2017), and various movement disorders (Hooi et al., 2017).

Moreover, neuroelectronic implants can restore lost sensory functions; for instance, retinal implants may return partial vision to those suffering from degenerative eye diseases, while cochlear implants restore hearing by directly stimulating the auditory nerve (Nella et al., 2022).

Brain-computer interfaces (BCIs) represent another transformative application of neuroelectronics. These devices can interpret neural signals into commands that control external devices, providing new ways for people with severe motor disabilities to communicate and move (Drakopoulou et al., 2023; Go et al., 2022).

Recent advancements in the field have also led to the creation of minimally invasive neural interfaces(Fanelli et al., 2022). These newer technologies, including soft, flexible, and injectable devices, enable less invasive implantations that reduce tissue damage and increase long-term biocompatibility, making them suitable for chronic applications (Seo et al., 2023).

In clinical settings, various neuroelectronic devices are already in use for both diagnostic and therapeutic purposes. Deep brain stimulation systems (Guimerà-Brunet et al., 2021) and vagus nerve stimulators (Adewole et al., 2019) are routinely used to manage conditions ranging from chronic pain to drug-resistant epilepsy (Ganguli et al., 2017) and severe depression (Holtzheimer et al., 2017).

### *1.2. Interfacing neurons with electronics*

Neuroelectronics, designed to interface with the nervous system, must be biocompatible to function safely and effectively within the body (Go et al., 2022). This ensures that these devices, whether implanted or in direct contact with neural tissues, do not trigger harmful biological reactions or adverse effects (Kireev and Offenhäusser, 2018). Central to neuroelectronics is the neuron-electrode interface, crucial for effective signal transduction between neurons and devices (Liang et al., 2021). The integrity of this interface is vital for the fidelity of neural signals, influenced by the quality of electrical contact and sealing resistance, essential for efficient stimulus transfer (Mariano et al., 2021; Tang-Schomer et al., 2014).

Biocompatibility in neuroelectronics involves careful material selection and engineering of surface, mechanical, and electrical properties to minimize immune responses and match the neural tissues' dynamic nature (Go et al., 2022). Enhancements in interface design utilize materials known for their excellent biocompatibility and adhesion (Milos et al., 2021), ensuring effective neuron-device integration. These include innovations like structured nano and microscale topographies that improve cell adhesion and alignment (Matino et al., 2020), critical for efficient neural stimulation and recording (Schiavone et al., 2020).

Furthermore, the interface's advanced features, such as electrochemical sensors for neurotransmitter detection (Reddy et al., 2019), offer invaluable insights for monitoring neurological conditions and guiding medical interventions (Kaur et al., 2022). These developments improve the clinical application of neuroelectronics and facilitate detailed studies of neuronal function, advancing our understanding of the nervous system and improving the long-term performance and safety of these devices (Keogh, 2020).

### *1.3. Key neuron features*

#### *1.3.1. Neuron anatomy*

The effectiveness of neuroelectronic devices is deeply tied to neuron anatomy, particularly through the neuron-electrode interface. This interface facilitates direct interaction between the complex structure of neurons and the functionalities of neuroelectronic devices. A thorough understanding of neuron anatomy is essential for improving device design and application in clinical and research settings. (Kandel et al., 2000).

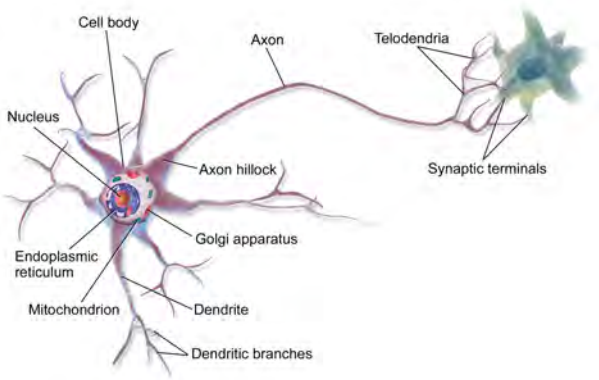



Figure 1: Anatomical structure of a neuron. Source: (Pitsis, 2018)

Neuron anatomy features several key components: the soma (cell body), dendrites, axons, and synaptic terminals (Figure 1).The soma houses the nucleus and is central to the neuron's metabolic and genetic activities. Dendrites extend from the soma and receive signals at synapses from other neurons. Axons not only carry impulses away from the soma to other neurons but also generate electrical signals that travel towards the synaptic terminals. At the axon ends, these impulses trigger the release of neurotransmitters, facilitating communication with other cells (Kandel et al., 2000).

Understanding the morphology and distribution of dendrites and axons is vital for precise placement of

electrodes. Electrodes positioned near dendrites can effectively capture incoming signals, while those near axons enhance the recording or initiation of outgoing signals. This strategic placement is crucial for optimal device function (Rinklin and Wolfrum, 2021).

### *1.3.2. Neuron polarity*

Neuronal polarity is a fundamental aspect of neuron structure and function, crucial for the effective communication and information processing within the nervous system. This polarity is characterized by the formation of a single axon and multiple dendrites from the neuron's cell body (Figure 2) . The axon transmits signals away, while dendrites receive incoming signals, establishing a directional flow of neural information as historically described by Santiago Ramón y Cajal, who noted that impulses typically travel from dendrites through the soma to the axon (Delgado-García, 2015).

Neuron polarity development is influenced by both positive and negative feedback loops. Positive feedback loops, which involve mechanisms that enhance the effects they initiate, significantly promotes axon growth in neurons (Zhou et al., 2020). Conversely, negative feedback loops work by initiating responses that reduce or inhibit their initial effects, help maintain neuronal polarity by restricting the formation of multiple axons, thus supporting dendritic growth (Takano et al., 2019). This polarity is vital during neurogenesis, where neurons form and orient themselves within existing neural circuits, and continues to play a role in axon guidance and synapse formation. Axons extend towards their targets guided by molecular cues, while synapses, forming primarily on dendrites, facilitate neuron-to-neuron communication.

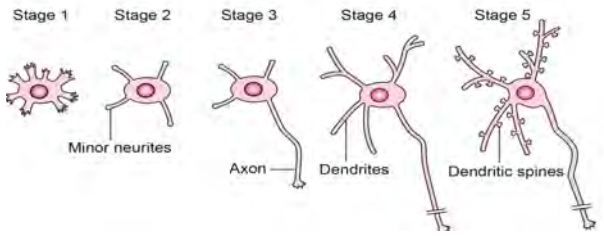



Figure 2: Stages of neuronal polarity development: initial symmetrical neuron progresses through neurite outgrowth, axon differentiation, dendritic branching, to mature synaptic spine formation.. Source: (Takano et al., 2019)

Through a series of well-defined stages depicted in Figure 2, neurons transform from multipolar precursors with equivalent neurites to functionally distinct structures. Initially, neurites undergo cycles of growth and retraction (Banker, 2018). One neurite then emerges as the axon, experiencing extended growth regulated by factors like microtubule dynamics (Higgs and Das, 2022). The remaining neurites differentiate into dendrites (Banker, 2018), establishing neuronal polarity. The specified axon continues to elongate and navigate towards its target, while dendrites mature for receiving synaptic inputs (Gärtner et al., 2015). This intricate process ultimately leads to the formation of functional neuronal circuits (Li et al., 2019).

### *1.3.3. Neuron communication*

Neuron communication is a fundamental aspect of nervous system function, unfolding primarily at synapses, where neurons meet (Batool et al., 2019). The role of neuronal polarity is critical in this process as it dictates the directionality of neural signals, ensuring precise integration with specific neuronal compartments (Gu et al., 2023).

Neuron communication is initiated by inputs from other neurons, which can lead to changes in the membrane potential of the neuron. This change can form an action potential, an electrical impulse if the inputs are strong enough to reach a threshold. Due to the neuron's polarized structure, this action potential then travels along the axon to the synapse. At the synapse, it triggers the opening of voltage-gated calcium channels, facilitating the influx of calcium ions into the neuron (Solecki, 2022). This influx is crucial as it prompts the release of neurotransmitters stored in synaptic vesicles (Kandel et al., 2000).

The neurotransmitters released into the synaptic cleft—the small gap between neurons—bind to specific receptors on the postsynaptic neuron (Szabo and Starke, 2021). This binding determines whether the postsynaptic neuron is more likely to fire its action potential (excitatory response) or less so (inhibitory response), depending on the types of neurotransmitters and receptors involved (Kandel et al., 2000). These inputs and their integration at the synapses dictate whether new action potentials will be formed, propagating the signal to the next neuron in the circuit.

Conversely, disruptions in these signaling pathways can manifest in various diseases. For instance, in multiple sclerosis (Bellingacci et al., 2021), the inappropriate activation of components within the signaling pathway leads to neuronal damage, while in depression (Parekh et al., 2022), altered chemical signaling due to dysfunctional glia-neuron interactions can affect overall brain function and mental health (Rudzki and Maes, 2021). These conditions highlight the importance of maintaining the structural and functional integrity of neurons.

### *1.3.4. Neuron structural components*

Neurons possess a complex cytoskeleton consisting of microtubules, actin microfilaments, and intermediate filaments, each crucial for maintaining neuronal structure and function. These elements not only provide stability but also allow for adaptations that are vital for neuronal health, offering potential targets for neuroelectronic interfaces to influence therapeutic morphological changes.

Microtubules, hollow structures about 25nm in diameter formed from tubulin dimers, extend from the neuronal cell body into axons and dendrites, facilitating the transport of vesicles and organelles crucial for neuronal function (Rafiq et al., 2022). Actin microfilaments, thinner at 4-6nm and formed by actin monomers, are concentrated in growth areas such as dendritic spines and near the plasma membrane, supporting neuronal growth, shape maintenance, and secretion processes (Shan et al., 2021).

Together, microtubules and actin filaments are essential for transporting synaptic vesicles and other organelles to synapses, enabling effective synaptic transmission and interneuronal signaling (Leshchyns' Ka and Sytnyk, 2016). Complementing them are intermediate filaments, or neurofilaments, 8-12nm in diameter, which form a matrix within axons that spaces microtubules and enhances the neuron's mechanical strength and axonal caliber (Rafiq et al., 2022).

Similar to the role of neuronal cell adhesion molecules (NCAMs) and L1 in facilitating neurite outgrowth, axon guidance, and synapse formation via their interactions with the cytoskeleton (Leshchyns' Ka and Sytnyk, 2016), integrins and paxillin function as a crucial mechanosensory and signaling unit for neuronal development. These transmembrane receptors (integrins) act as extracellular tethers, binding the neuron to the surrounding extracellular matrix (ECM) (Bokel and Brown, 2002). Internally, paxillin serves as a critical adaptor protein within focal adhesions, bridging the gap between integrins and the actin cytoskeleton (López-Colomé et al., 2017). This intricate interplay allows neurons to not only maintain their structural integrity and facilitate intracellular transport but also dynamically respond to environmental cues.

### *1.4. Methods to understand neuron behavior*

To thoroughly understand neuron behavior, researchers utilize a range of methods that dissect the complex interactions and structures of neurons. A fundamental aspect of this research is electrophysiology, which involves techniques like intracellular recordings and stimulation using sharp or patch electrodes to measure ionic currents and voltages within individual neurons precisely (Paternò et al., 2021). Extracellular recordings with microelectrode arrays (MEAs) offer a non-invasive approach for monitoring multiple neurons simultaneously, crucial for analyzing network activity (Hales et al., 2010). However, the effectiveness of MEAs heavily relies on their design and patterning.

Precise electrode size, shape, and arrangement are crucial for recording or stimulating specific neuronal populations (Viswam et al., 2019). Denser electrode arrays with smaller features enable recording activity from individual neurons (Muthmann et al., 2015), while elongated finger-like electrodes can target specific subcellular compartments like axons or dendrites (und Halbach, 2009). Microelectrode patterning techniques like photolithography offer control over these features, ensuring the MEA effectively interacts with neurons for the desired study (Temiz et al., 2012).

Fluorescence microscopy extends its utility by assessing biocompatibility. By monitoring the health and behavior of fluorescently labeled cells grown on the microelectrodes (Khan et al., 2011), researchers can indirectly assess if the electrodes cause any harm to the neurons.

Advanced microscopy techniques like total internal reflection fluorescence microscope (TIRFM) and super-resolution microscopy offer high-resolution imaging of live and fixed neurons, revealing details of neuronal structure and dynamics (Rossi et al., 2018). TIRFM provides high-contrast imaging for studying live neuronal cultures and axonal dynamics (Opstad et al., 2020), while super-resolution microscopy allows detailed visualization of synaptic protein localization (Nosov et al., 2020). Fluorescence correlation spectroscopy can quantify protein dynamics in live neurons by tracking fluorescently labeled proteins (Fujita et al., 2020).

Furthermore, brightfield microscopy complements fluorescence techniques by offering label-free visualization. It can be combined with fluorescence imaging to correlate neuronal morphology with molecular localization patterns, providing a comprehensive view of neuronal structure and function (Schmued et al., 1989).

These optical imaging techniques, combined with electrophysiological methods, are critical for developing a comprehensive view of neuronal behavior, particularly in how neurons adapt their responses and morphology under varying conditions, enhancing the study of synaptic plasticity, neurotransmitter dynamics, and overall neuronal health (Claverol-Tinture et al., 2005).

Despite significant advancements in neuroelectronic and imaging technologies, there remains a considerable gap in accurately characterizing and analyzing neuronal morphology and polarity, which are intrinsically unpredictable due to their complex biological nature. This unpredictability poses unique challenges for the current algorithms and methods employed to study neuronal behavior and structure.

## 2. State of the art in automated image analysis in cellular and neuronal research

The progression of image analysis techniques in neuronal studies has been marked by significant advances, beginning with tools like the Neuron Image Analyzer (NIA). This technology enhances neurite tracing and structural identification using methods such as the Laplacian of Gaussian (LoG) filter and Level Set Method (LSM), which notably reduce the reliance on manual annotation, previously facilitated by tools like

Nikon's NIS Elements. Supported by MATLAB, NIA represents a shift from manual to automated processes, significantly improving precision and reducing the labor intensity traditionally required in neuronal studies (Kim et al., 2015).

Building upon these foundational techniques, CellProfiler introduces a modular pipeline that greatly facilitates image analysis, ranging from image loading to detailed object measurement. This flexibility is particularly crucial for handling the diverse imaging requirements inherent in neuronal research, seamlessly integrating with other powerful tools such as Ilastik and ImageJ (Lamprecht et al., 2007). Ilastik specializes in machine learning-based image segmentation, handling complex textures, and integrating with CellProfiler for efficient batch processing of large datasets (Sommer et al., 2011). Complementarily, ImageJ, enhanced by the NeuronJ plugin, offers robust, open-source image processing across various systems, making it an ideal choice for neuron-specific analyses (Abràmoff et al., 2004).

A practical application of these integrated tools is demonstrated in the work of Ossinger et al. (2020), which employs ImageJ, Ilastik, and CellProfiler to analyze brightfield micrographs. These tools adeptly handle the challenges posed by uneven backgrounds and variable intensity, with ImageJ's adaptive thresholding techniques assessing axonal outgrowth, while Ilastik and CellProfiler quantify dendritic development. This synergy showcases the efficacy of these tools in providing detailed and accurate neuron analysis, marking a significant step towards automated and precise characterizations of neuron morphology.

Advancing into the realm of deep learning, the NeuroCyto platform addresses the challenge of neurite crossover with a directed graph model, enhancing neurite tracing and effectively separating crossed neurites using dual-channel imaging. This platform is supported by advanced algorithms that automate the segmentation and analysis of fluorescence images, thereby enabling the extraction of quantitative parameters crucial for comprehensive cellular analyses (Schurr et al., 2023). Following this, DeepNeuron extends these capabilities into 3D neuron tracing, employing convolutional neural networks (CNNs) for sophisticated foreground/background classification, which facilitates the detection of neurite signals without prior preprocessing. Additionally, its revised Siamese network aids in connecting neurite structures from detected signals, further refining neuron morphology by filtering out false positives (Zhou et al., 2018).

The innovative NeuriTES platform leverages adaptive semantic segmentation for tracing motor neuron evolution over time in bright-field time-lapse microscopy, specifically targeting neuron degeneration in ALS studies. This approach avoids the pitfalls of phototoxicity and interference associated with fluorescent labeling, with initial frames manually labeled using ImageJ to train the segmentation network, ensuring accurate neuron identification throughout the study (Mencattini et al., 2021). Complementing this, instance segmentation models like Mask-RCNN generalize across different imaging modalities by automatically detecting and contouring neuron somas from fluorescence microscopy images (Tong et al., 2021), showcasing the adaptability of these models to various experimental conditions.

Furthermore, the comprehensive framework proposed by Mari et al. (2015) for the quantitative and morphological analysis of rat dorsal root ganglion neurons cultured on MEAs provides detailed metrics such as neuron-to-neuron and neuron-to-microelectrode distances, offering invaluable insights into the organization and dynamics of neuronal networks. The study focuses on the segmentation of neurons from fluorescence channel images using thresholding, watershed transform (Vincent and Soille, 1991), and object classification (Liu et al., 2021), while microelectrode positions are identified from transmitted light channel images via the circular Hough transform (Illingworth and Kittler, 1987).

In a broader context, the study by de Santos-Sierra et al. (2014) investigates the development of small-world network configurations in vitro cultures of dissociated invertebrate neurons from locust ganglia. Utilizing custom image analysis software, this research tracks the self-organization of these cultures into complex networks characterized by high clustering and short path lengths, indicative of efficient neuronal processing and network resilience. Such insights are critical for understanding the network dynamics and the morphological changes throughout the development stages.

Moreover, the studies by Radotić et al. (2017) and Onesto et al. (2019) explore how microelectrode arrays and nanowire substrates influence the alignment, orientation, and assembly of neuronal cells into functional networks. These studies highlight the significant impact of substrate topography on neuronal behavior and network connectivity, crucial for applications in neural tissue engineering and understanding cortical-like minicolumns.

As the field progresses, addressing the challenges posed by background noise, photobleaching, and substrate variability will require optimized imaging protocols and advanced signal-processing algorithms. The adaptability and refinement of deep learning models like Convolutional Neural Networks (CNNs) (O'shea and Nash, 2015), U-Net (Yin et al., 2022), and Mask-RCNN (Region-Based Convolutional Neural Networks) (He et al., 2017) are essential for maintaining accuracy and reliability, paving the way for novel methodologies in neuronal behavior studies and expanding the potential for future research endeavors.

## 3. Aim of the work: enhancing analytical methodologies in neuronal imaging

While micro-electrode arrays and optical imaging technologies provide essential insights into neuronal behavior and morphology, they encounter significant limitations during the data analysis phase (Luan et al., 2023). Current analytical tools struggle with accurately identifying neuron boundaries, a crucial task complicated by the intricate and varied morphology of neurons (Mencattini et al., 2021). These challenges often result in frequent errors in automated segmentation algorithms and necessitate labor-intensive manual or semi-automated methods to ensure accuracy. This reliance on time-consuming correction processes significantly impedes the efficiency and scalability of neuronal research (Al-Kofahi et al., 2006).

The complexity of neuronal structures, combined with their unpredictable growth patterns and interactions, underscores the need for more sophisticated analytical methodologies capable of handling this variability with high precision (Friedrich et al., 2013). Addressing these challenges is crucial for advancing our understanding of neuronal function and disorders, leveraging the full potential of neuroelectronic and imaging technologies.

## 4. Project goals

Our project aims to refine and enhance the methodologies used in neuronal imaging and analysis to gain deeper insights into neuronal behavior. Given the novelty and complexity of our dataset, our focus will be on developing exploratory advancements to address these challenges:

1. Advanced Neuron Morphology Characterization:
   Refine image analysis techniques for detailing neuron morphology, including behavior clustering, size measurements, and structural tracking.
2. Quantification of Protein Expression:
   Implement methods to analyze fluorescence intensity across imaging channels, providing insights into neuron function and responses to environmental changes mediated by MEAs.
3. Network Characterization:
   Apply advanced segmentation and graph analysis to better understand cortical neuron nuclei connections and network dynamics.

Addressing these goals requires leveraging the latest advancements in automated image analysis tools and deep learning methodologies. The subsequent section reviews the state of the art in these areas, focusing on both the achievements and the challenges that need to be overcome to meet our project objectives.

## 5. Material and methods

### *5.1. Datasets*

These datasets, each utilizing distinct imaging techniques and biological markers, are crucial for developing and testing algorithms that analyze cellular behaviors and interactions under varying conditions.

For cluster identification and network characterization, the dataset consists of fluorescent images of mouse cortical cell nuclei, it indirectly tackles neuronal morphology through cluster formation. Actin-MAP2 hints at neurite presence, a key morphological aspect, while Phosphorylated Paxillin (p-Pax)-Pax-Tau1 sheds light on actin dynamics potentially affecting clustering.

The dataset is unique because the cells are cultured on microchip substrates with specifically designed micropillar arrays. These micropillars vary in diameter (thickness) and pitch (spacing between pillars). This intricate topography allows us to investigate how physical cues from the environment, governed by micropillar characteristics, influence neuronal organization.

For soma and neurite identification, the dataset features labeled neurons with highlighted somas and neurites, positioned on different structural substrates like stubby, mushroom, and thin formations. This dataset facilitates the identification and analysis of soma and neurite structures, crucial for understanding neuron morphology across varied topographical contexts.

A live imaging dataset, this dataset utilizes bright-field microscopy, which effectively highlights the different grid patterns of the micropillar arrays. By capturing real-time neurite growth and behavior on these varying topographies, researchers can directly track neurite extension and assess how the physical structures influence neuronal morphology. This offers valuable insights into developmental processes and how neurons adapt to their environment.

Additionally, for protein expression analysis, the dataset utilizes fluorescent images labeled for key components of focal adhesions, structures crucial for anchoring neurons and influencing polarity. The dataset focuses on chicken embryo cortical cells, it presents a unique challenge for analyzing neuronal structure due to the presence of fluorescent pillars. While these pillars offer valuable information about the substrate topography, their brightness can interfere with the distinction of neuronal structures, which are crucial for protein expression analysis. These pillars are not uniform; they vary in shape—thin, mushroom, and stubby—and are organized at different pitches (p10, p4, p30), including a control flat environment. This variety allows the examination of how different physical substrates influence neuronal behavior and development.

### *5.2. Methodology*

#### *5.2.1. Milestone 1.1: Nuclei and clusters identification*

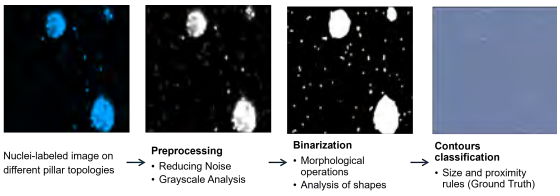



Figure 3: Image Analysis to identify cluster in images

1. Pseudo-labeling: The primary objective of this pipeline is to generate pseudo-labels for training a Mask-RCNN model (Figure 3), which is used to identify and segment clusters of nuclei.
   Initially, contrast stretching is applied to maximize the image's dynamic range. Subsequently, binary thresholding simplifies the image to its fundamental shapes, facilitating the isolation of key features through a combination of morphological operations such as closing, eroding, and dilating. If initial thresholding is inadequate, additional preprocessing such as edge detection using the Sobel operator generates the binary mask.
   The adjusted images are then analyzed to detect and classify contours using OpenCV. Contours represent potential nuclei and are classified based on their area and solidity into small (single nuclei) or potential clusters.
   For contours identified as potential clusters, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm is employed to spatially cluster contour points, distinguishing between closely packed nuclei clusters and individual nuclei or noise-based on density.

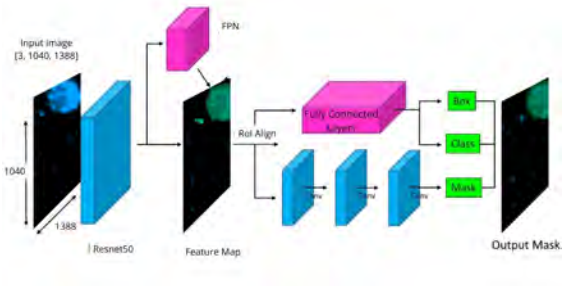



Figure 4: Image analysis to identify cluster in images

2. Mask-RCNN training and validation: The pipeline (Figure 4) initiates by configuring a Mask-RCNN model pre-trained on the COCO dataset, leveraging a ResNet-50 backbone equipped with an FPN for efficient feature extraction across scales. The original heads of the model, responsible for classification and mask prediction, are replaced to accommodate the specific requirements of the task—identifying background versus nuclei clusters. These new heads are initialized using He initialization to ensure that the model trains effectively from the start.
   The dataset is prepared for both training and validation, these transformations include random horizontal flipping, normalization, and conversion to tensors.
   Training is conducted using an SGD optimizer with parameters like learning rate, momentum, and weight decay to govern the learning process. A learning rate scheduler is used. During each epoch, the model is updated by computing and back-propagating losses, including classification, bounding box regression, and mask prediction.
   For evaluation, the Intersection over Union (IoU) metric is used for quantifying the accuracy of the model's predictions. IoU is computed for both bounding boxes and masks, offering a detailed measure of how well the predicted outputs align with the pseudo-labels.

### *5.2.2. Milestone 1.2: Soma and neurite identification*

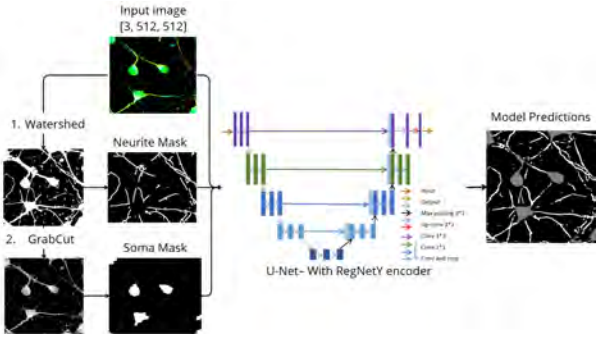



Figure 5: Soma and neurite identification pipeline

The main idea of this pipeline (Figure 5) is to use a sequence of image processing techniques—ranging from preprocessing and segmentation to soma and neurite identification, followed by association analysis—to generate detailed pseudo-masks for training a U-Net model, thereby enhancing the accuracy and efficiency of identifying and labeling neuronal structures in images.

1. Identifying of Soma and Neurites
   (a) Preprocessing: The process begins converting the image to grayscale. To reduce image noise, a non-local means denoising technique is applied. Following denoising, thresholding using Otsu's method (Yousefi, 2011) is applied to create a binary mask that separates potential areas of interest from the background. Then, Contrast-Limited Adaptive Histogram Equalization (CLAHE) (Reza, 2004) is applied.
   (b) Watershed Algorithm: The binary mask is created through thresholding. Then, a distance transform is made, which computes

for each foreground pixel the nearest distance to the background, creating a relief map where peaks correspond to the centers of objects. Local maxima of this map are identified as markers—potential starting points for the Watershed algorithm (Vincent and Soille, 1991). The watershed then treats these markers and defines the boundaries of separate objects.

(c) GrabCut Algorithm (Wang et al., 2023): Starting with the initial segmentation provided by the Watershed algorithm, the image and a mask delineating probable and definite regions of foreground and background are prepared. Morphological operations further define definite areas—erosion shrinks the foreground regions to ensure only the most certain parts remain, while dilation expands the background regions, helping to exclude less certain border areas.

(d) Soma Identification: The process involves analyzing the contours extracted from the refined segmentation mask of the GrabCut step. Each contour is evaluated based on its area, perimeter, circularity, and solidity to determine if it matches the characteristics typical of somas. Contours that meet these criteria are filled in on a new mask specific to somas.

(e) Neurite Identification: This step starts by converting the watershed-labeled image to a binary mask that excludes soma regions. The masks are then skeletonized (Abu-Ain et al., 2013). The skeleton undergoes a cleaning process to remove fragments shorter than a specified branch length and is enhanced through morphological operations to improve connectivity between segments.

2. U-Net training and validation

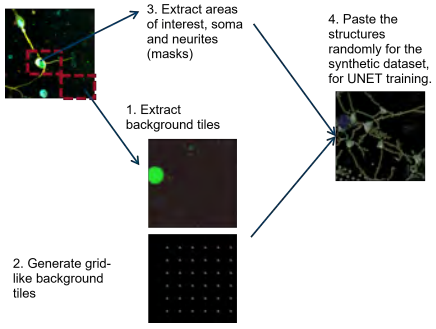



Figure 6: Synthetic data generation pipeline

(a) To enhance the model's exposure to varied backgrounds and increase the robustness of our predictions, the new synthetic dataset was synthesized by extracting key structures from original images and pasting them onto different backgrounds ( Figure 6), both with and without noise. This approach allowed the simulation of more diverse imaging conditions, preparing the model for real-world applications where background variability can be significant.

(b) The efficacy of the U-Net model (Roy et al., 2018) was tested with different encoder backbones including RegNetY320, VGG16, MobileNetV2, EfficientNet-B7, and ResNet152. The RegNetY-320 model emerged as the most effective. This was possible thanks to the segmentation models PyTorch framework (SMP) (Iakubovskii, 2019). The RegNetY-320 model(Radosavovic et al., 2020) is then used as the encoder. RegNetY is a convolutional network design space with simple, regular models with parameters: depth, initial width, and slope, and generates a different block width for each block. It has 141.3 million parameters and the capability to process high-dimensional data efficiently, thanks to techniques like stochastic depth, gradient checkpointing, and layer-wise learning rate decay. These features help manage the computational load and optimize training dynamics.

(c) Given the limited size of the real dataset, a K-Fold cross-validation approach is employed to maximize the usage of available data for both training and validation. This method divides the dataset into 'k' subsets, and iteratively, each subset is used for validation while the others are used for training.
During training, the model uses a combination of Jaccard, Dice, and Focal loss functions to handle the class imbalance and enhance the learning of relevant features. These loss functions are weighted to adjust their impact on the overall training process, optimizing the model to improve overlap (Jaccard and Dice) and focus more on difficult-to-classify pixels (Focal), like the neurites.

(d) Evaluation: Metrics such as Intersection over Union (IoU), Dice scores, and F1 recall are monitored for training and validation to gauge the model's performance.

(e) Post-processing of mask for analysis: In the postprocessing of neurite-soma masks, binary dilation is first applied to soma masks using a disk-shaped footprint to expand their areas, enhancing the likelihood of neurite-soma intersection for nearby neurites. Subsequently, both soma and neurite masks undergo connected component labeling to identify individual regions. Next, the dilated soma mask is overlayed onto the neurite mask to ascertain which soma region over-

laps most significantly with each neurite, determining their associations based on pixel overlap frequencies. The quantitative metrics—normalized neurite length and number—are crucial for standardizing neurite assessments relative to cell counts, ensuring that results are comparable across samples and conditions (Figure 17b). These metrics provide a robust framework for evaluating neurite proliferation and extension, which are indicative of neuronal health and network capabilities. Normalized neurite length is given by:

$$L_{norm} = \frac{L_{total}}{N_{cells}}$$

where $L_{norm}$ is the normalized neurite length, $L_{total}$ is the total neurite length, and $N_{cells}$ is the total number of cells. Additionally, the normalized neurites number is given by:

$$N_{norm} = \frac{N_{neurites}}{N_{cells}}$$

where $N_{norm}$ is the normalized neurite number, $N_{neurites}$ is the total number of neurites.

### 5.2.3. Milestone 1.3: Neurite tracking

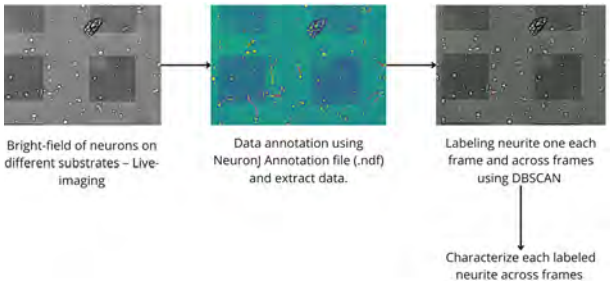



Figure 7: Neurite tracking pipeline using DBSCAN algorithm

1. Tracing Neurite Coordinates: The initial step involves converting NeuronJ tracing data (Figure 7), which consists of vertices representing neurite endpoints, into continuous coordinate paths. By connecting these vertices, the script recreates the full trajectory of each neurite within the image frames. This process ensures that each neurite is represented as a continuous entity, facilitating subsequent analyses that rely on tracing entire neurite paths.
2. Intra-Frame DBSCAN Clustering: Within each frame, the DBSCAN algorithm clusters the neurite coordinates to identify distinct neurite entities. This clustering helps differentiate individual neurites based on the density of traced points. For each cluster, the script calculates the centroid, serving as a representative central point, and the total length of the neurite, which is derived from the sum of distances between connected coordinate points.
3. Inter-Frame DBSCAN Clustering: After identifying neurites within individual frames, a second DBSCAN clustering is applied across all frames using the centroids from the first clustering stage. This inter-frame analysis aims to track the persistence and evolution of neurites over time by identifying clusters that appear consistently across the dataset. This step is crucial for monitoring dynamic changes and behaviors in neurite structures through sequential imaging data.
4. Mask-RCNN training and evaluation for detecting neurites:

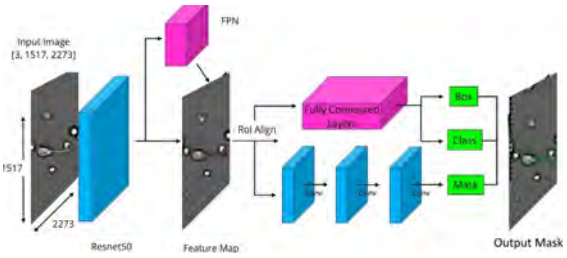



Figure 8: Mask-RCNN model to Identify neurites in Live Imaging

The dataset is structured into tiles, each representing a segment of video data containing detailed neurite images. Each neurite is annotated with its bounding box and mask, facilitating precise instance segmentation. Transformations applied to the dataset include photometric distortions, random zoom-outs, and horizontal flips to enhance model robustness by simulating various imaging conditions.

The Mask-RCNN model, specifically configured with a ResNet-50 backbone and FPN, is employed for neurite segmentation (Figure 8). The training process utilizes SGD optimizer, as outlined in the original Mask-RCNN paper, with specific parameters like learning rate, momentum, and weight decay. The learning rate scheduler adjusts the rate during training to optimize performance. Model outputs include class predictions, bounding boxes, and segmentation masks, which are iteratively refined through the epochs.

In the evaluation phase, the Intersection over Union (IoU) metric is calculated for both bounding boxes and masks to assess model accuracy. This involves matching predicted boxes and masks to their respective ground truth annotations based on IoU scores, which provides a quantitative measure of model precision in segmenting neurites accurately.

### 5.2.4. Milestone 2: Protein expression

Images are loaded and processed to extract individual channels (Figure 9). For each region (soma, neurite, and

the compound of both), masks are applied to the original image to segment the area of interest. The mean intensity for cells and background is calculated within these masks. Normalized fluorescence intensity is computed by subtracting the background mean intensity from the cell mean intensity and normalizing by the area of the mask, as the following equation shows:

$$P_e = \frac{\bar{N}_I - \bar{B}_I}{N_{\text{area}}}$$

where $P_e$ is the protein expression, $\bar{N}_I$ is the neuron mean intensity, $\bar{B}_I$ is the background mean intensity, and $N_{\text{area}}$ is the neuron area.

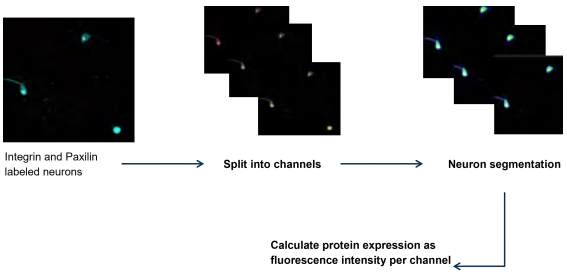



Figure 9: Protein Expression analysis Pipeline

### 5.2.5. *Milestone 3: Network analysis of nuclei*

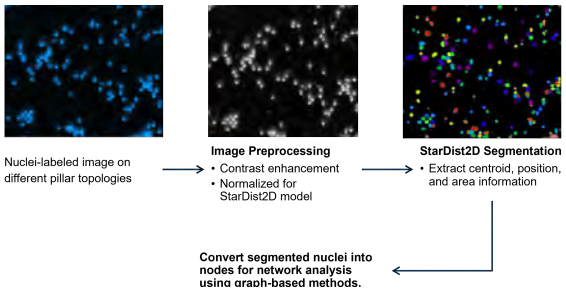



Figure 10: Nuclei Segmentation and Graph-based Network Analysis Pipeline

This pipeline (Figure 10) employs the StarDist2D model (Schmidt et al., 2018; Weigert and Schmidt, 2022), a U-Net-based neural network specialized in identifying single nuclei from biological images. It was necessary to clean up data to ensure only single nuclei were analyzed, excluding larger clusters that may be inaccurately identified as single entities.

Before segmentation, the pipeline enhances image contrast and saturation using CLAHE and color adjustments to make the features more distinct and easier for the model to identify. The StarDist2D model segments the processed images, identifying potential nuclei based on shape and intensity.

A Waxman (Waxman, 1988) graph is initially created based on the centroids of the segmented nuclei. The Waxman model is specifically made for routing of multipoint connections and generates a network structure that (such as the structural brain networks) allows information transfer across the network nodes (Onesto et al., 2019). This graph randomly connects nodes with a probability decreasing with the Euclidean distance between them, scaled by parameters alpha and beta. The graph is enhanced by adding edges based on KNN, where edges between the nearest neighbors are weighted by their distances.

A cluster Analysis is also carried out using the clustering algorithm, DBSCAN, on the node positions to further analyze spatial distributions and cluster formations among the nuclei.

## 6. Results

### 6.1. *Milestone 1.1: Nuclei and cluster identification*

The successful application of image analysis techniques yielded visually accurate segmentation results to later use to train the Mask-RCNN model (Figure 11). This figure confirms the pipeline's capability to identify clusters and single nuclei on different substrates. Such high-resolution differentiation is critical in studies where cellular behavior in response to microenvironmental features is analyzed.

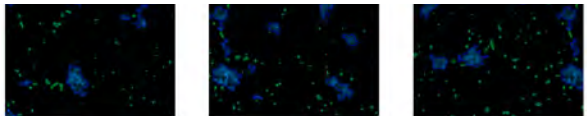


Figure 11: Identification of clusters (blue contours) and single nuclei (green contours) on the specific micro-patterned substrate of pillars with a diameter of 2 micrometers and a pitch of 20.

Figure 12 provides a visual representation of the distribution of clusters across different micro-patterned substrates. This analysis shows how physical microenvironmental cues can influence cellular organization and cluster formation. The ability to visualize and quantify this distribution allows to draw correlations between substrate patterning and biological outcomes.

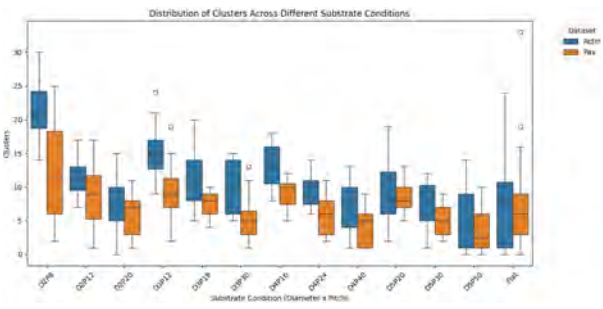



Figure 12: Distribution of clusters across the different micro-patterned substrates of pillars

As shown in Figure 13, the loss metrics during the training of the Mask-RCNN indicate a stable and converging training process over 20 epochs with Xavier weight initialization. The consistent reduction in loss values reflects effective learning and adaptation by the neural network to the task-specific features of nuclei segmentation.

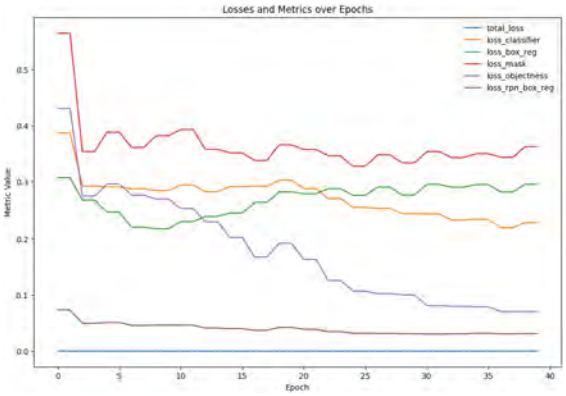



Figure 13: Loss evolution during Mask-RCNN training (20 epochs) with Xavier weight initialization

The Intersection over Union (IoU) metrics—0.901 for boxes and 0.852 for masks—highlight the model's high accuracy in detecting and segmenting nuclei as demonstrated in Figure 14. These metrics underscore the model's high accuracy in not only detecting the correct location of the nuclei (bounding boxes) but also in accurately segmenting (masks) the nuclei from the background.

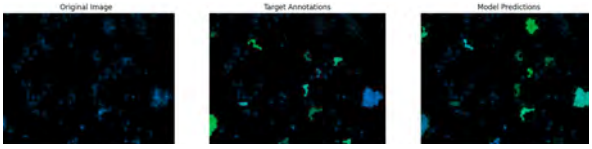



Figure 14: Cluster Identification by Mask-RCNN training with Xavier weights initialization

### *6.2. Milestone 1.2: Soma and neurite identification*

In the ongoing effort to enhance the accuracy and reliability of soma and neurite identification, various U-Net models equipped with different backbone architectures were evaluated. The results, summarized in the table below, showcase the performance metrics obtained after 10 epochs of training.

| Backbone | Loss | IoU | F1 |
|---|---|---|---|
| RegNetY | 0.2398 | 0.8634 | 0.9266 |
| VGG16 | 0.2434 | 0.8603 | 0.9248 |
| MobileNET | 0.2605 | 0.8500 | 0.9188 |
| EfficientNet | 0.2726 | 0.8379 | 0.9117 |
| ResNet 132 | 0.2468 | 0.8584 | 0.9237 |

Table 1: Accuracy results after training (10 epochs) the U-Net with different backbones

RegNetY stands out with the highest F1-score and IoU, indicating superior segmentation capability, particularly effective in delineating complex neuronal structures. Although all models demonstrated high competence, the gradient in performance metrics from RegNetY to EfficientNet highlights the influence of network architecture on segmentation tasks. Smaller models like MobileNET, while efficient, offer slightly reduced accuracy.

The graphical representation in the figure 15 captures the evolution of loss and Intersection over Union (IoU) scores during a K-fold cross-validation training regimen. The training utilized the advanced RegNetY-320 model, selected for its robust architectural benefits conducive to handling complex neuronal structures.

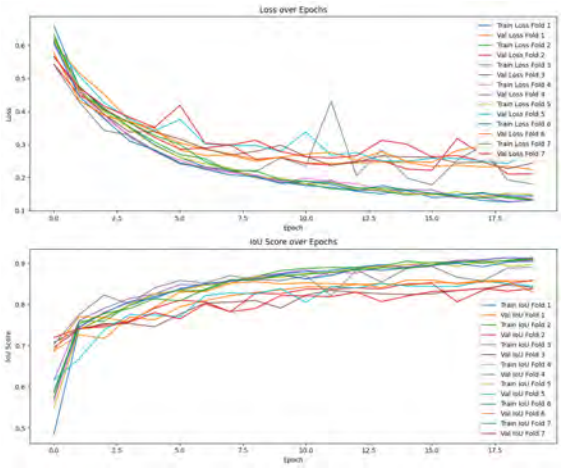



Figure 15: Training and Validation Metrics for U-Net Model Across Different Folds. Top graph displays the loss metrics over epochs, highlighting the decreasing trend in training and validation loss across seven-folds. The bottom graph shows the Intersection over Union (IoU) scores, where both training and validation IoU gradually increase.

The effectiveness of the RegNetY backbone is visually confirmed in figure 16, illustrating the precision in predicting soma (gray) and neurites (white). The comparison between the pseudo-mask (true mask) and the predicted mask validates the high performance of the model. It confirms its utility for detailed morphometric analyses and quantitative assessments of neuronal structures.

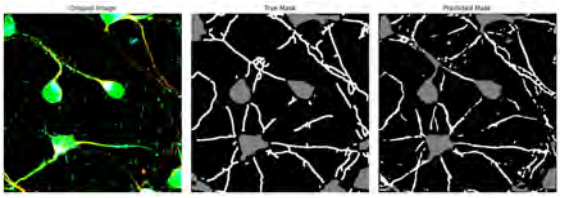
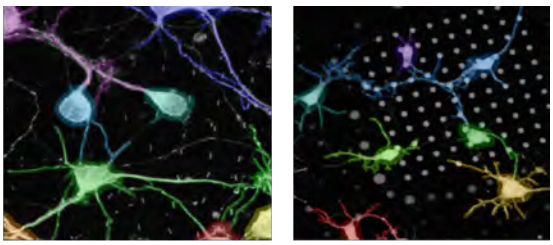
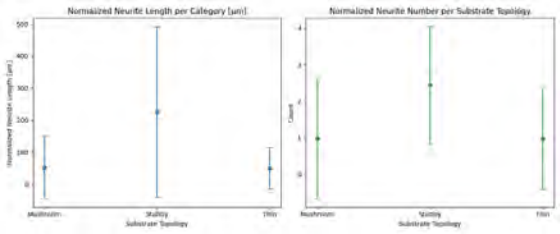


Figure 16: Pretrained U-Net with RegNetY backbone prediction of soma (gray) and neurites (white) at Fold 3, Epoch 20.

Post-processing techniques have further refined the segmentation accuracy, as shown in the sub-figures which present post-processed masks (Figure 17a). By employing binary dilation and connected component labeling, the overlap between soma and neurite regions is enhanced, allowing for precise measurement of neurite lengths and counts relative to the number of cells. This approach ensures that neurite assessments are standardized and comparable across different experimental conditions, providing a robust framework for evaluating neurite proliferation and extension (Figure 17b).



(a) Post-Processing of soma and neurite mask results



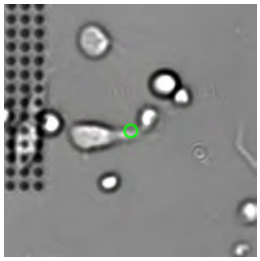

(b) Neurite analysis (normalized length and number) on different topologies

Figure 17: Results from soma and neurites identification to study neuronal responses to different topologies

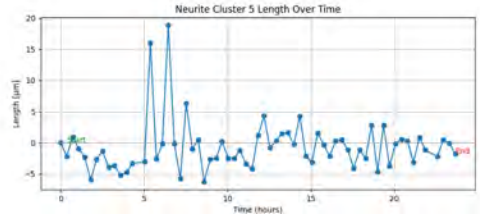
### *6.3. Milestone 1.3: Neurite tracking*

The pipeline for neurite tracking, depicted in Figure 7, employs image processing and clustering techniques to track neurites over time. This process leverages the DBSCAN algorithm to cluster neurite coordinates within and across frames and uses the Mask-RCNN model to segment neurites in live imaging.

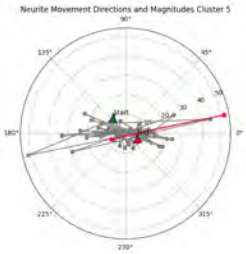
The conversion of NeuronJ tracing data into continuous coordinate paths allows for the representation of each neurite as a continuous entity. It allowed for tracking the detailed morphology and dynamics of neurite growth. DBSCAN clustering within frames identifies distinct neurite entities. Subsequent inter-frame clustering tracks these neurites over time, providing insights into their temporal stability and morphological changes. Figures 18 illustrate the initial state, length changes, directional movements, and growth velocity of a neurite across time frames. It provides a comprehensive view of neurite behavior over time.



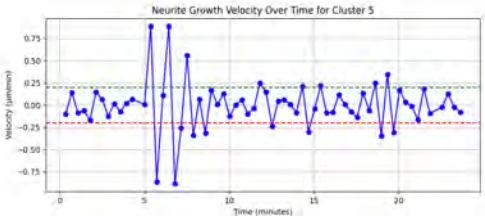
(a) Initial frame of a sequence capturing a labeled neurite (Neurite no. 5), highlighted using a green marker. Baseline reference for tracking subsequent growth and directional movements of the neurite, establishing the starting point for detailed analysis.




(b) Tracking length changes across frames of neurite. Variations in the length of the neurite over time, measured across multiple frames. It indicates periods of growth and retraction, which are essential for understanding the underlying biological processes affecting neurite dynamics.




(c) Direction of displacement of identified neurite. It visualizes the directional displacement of the neurite throughout the observation period. The red line indicates the predominant direction of neurite extension.

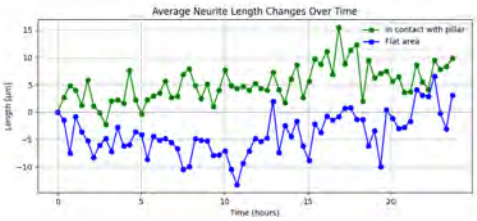



(d) Growth velocity of the neurite ($\mu$m per minute), quantified in micrometers per minute. It highlights phases of advancing (green line) and retracting (red line) growth.

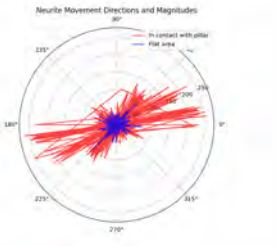
Figure 18: Neurite tracking analysis of a single neurite sample (Neurite no. 5)

Furthermore, it is also possible to study the behavior of all neurites like in Fig. 19 whether the interaction

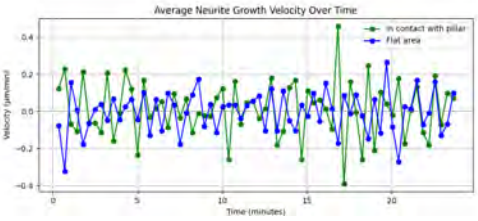
of pillars affects the neurite dynamic changes. These plots delineate the average length changes and growth velocities of neurites over time, contrasting neurites in contact with pillars to those in flat areas.




(a) Tracking average length changes across frames of all labeled neurites in terms of pillar interaction




(b) Direction of displacement of all identified neurites in terms of pillar interaction




(c) Velocity (micrometer per minutes) of all identified neurites depending on its interaction with pillars

Figure 19: Neurite tracking analysis of labeled neurites

The application of Mask-RCNN facilitates high-precision neurite segmentation. This model's ability to discern neurites against complex backgrounds is crucial for accurate tracking and analysis. The training of the Mask-RCNN model shows a progression in loss reduction, as indicated in figure 20. However, the average IoU scores for boxes (0.5756) and masks (0.2692) suggest moderate segmentation accuracy, indicating potential areas for model refinement or training data enhancement.

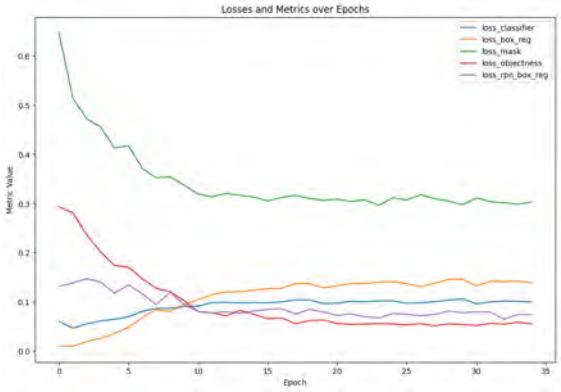



Figure 20: Loss evolution during training

### 6.4. *Milestone 2: Protein expression*

The approach allows for a targeted evaluation of protein localization and concentration by utilizing fluorescent labeling of proteins such as Paxilin and Integrin, crucial for understanding cellular adhesion mechanisms.

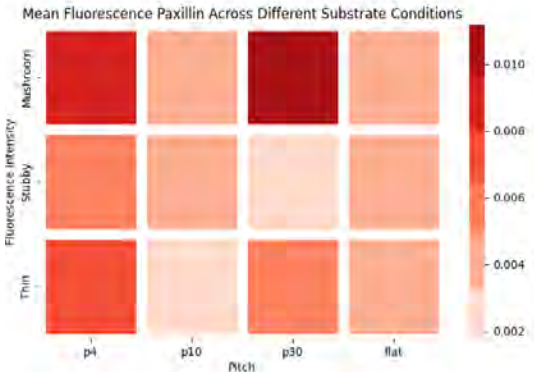



Figure 21: Protein expression heatmap measurement of Red Channel or Paxilin channel data on different substrate topology against flat surface.

The heatmap, shown in figure 21, measures the expression of Paxilin across various substrate topologies compared to a flat surface. This channel specifically highlights the areas where Paxilin—a protein involved in focal adhesion—is most concentrated, indicating regions of active cellular engagement and structural anchoring.

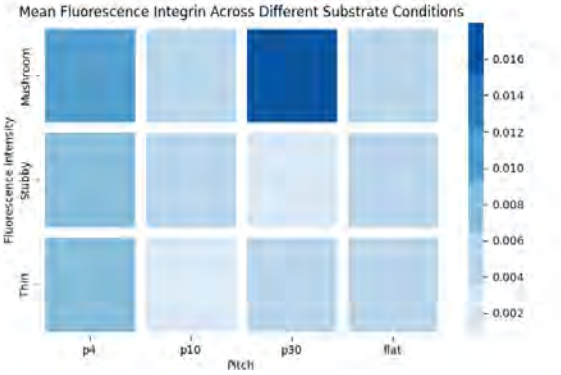



Figure 22: Protein expression heatmap measurement of Blue Channel or Integrin channel data on different substrate topology against flat surface.

Similarly, figure 22 measures the expression of Integrin, another key protein in cell adhesion processes. The heatmap provides a comparative analysis against different substrate topologies, underscoring how topographical variations can influence Integrin distribution and, by extension, cell adhesion dynamics.

### *6.5. Milestone 3: Network analysis of nuclei*

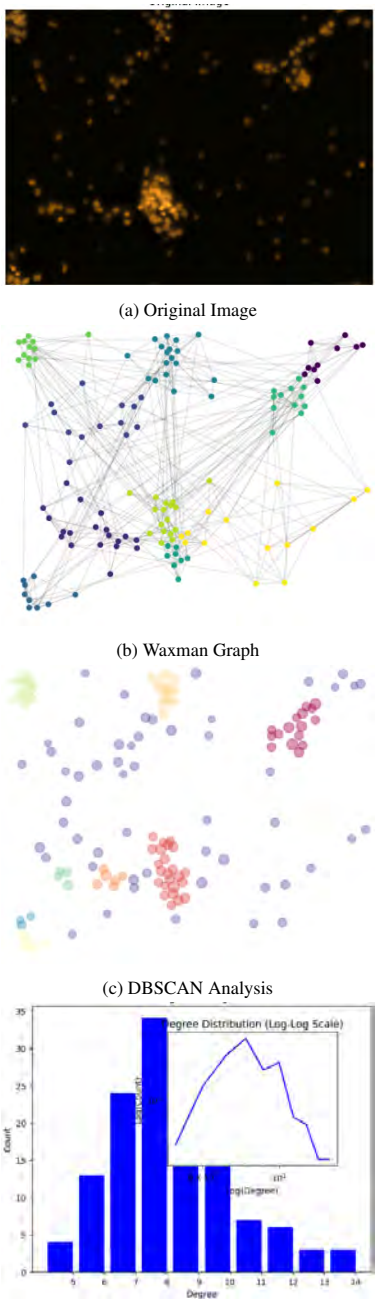
The analysis pipeline, as illustrated in Figure 10, utilizes the StarDist2D model for precise nuclei segmentation from biological images. This model is specifically adapted to detect individual nuclei by enhancing image contrasts and removing clusters that may be misidentified. The segmentation results feed into a graph-based analysis to explore the spatial relationships and clustering of nuclei.

Post-image enhancement via CLAHE and color adjustments (Figure 23a), the StarDist2D model successfully segments the nuclei, emphasizing distinctiveness in shape and intensity crucial for accurate identification.

From the segmented data, a Waxman graph is constructed using nuclei centroids (Figure 23b). This graph models the probability of connections between nodes (nuclei) inversely related to their Euclidean distances, adjusted by the alpha and beta parameters to optimize the network structure. In addition to the Waxman model, a K-nearest neighbor (KNN) approach is applied, introducing edges weighted by the physical proximity of the nuclei, which enriches the graph's connectivity and relevance to actual biological structures (Figure 23c).

DBSCAN (Ester et al., 1996) is employed to analyze the spatial distribution and cluster formation among the nuclei (Figure 23c). This method helps in identifying and analyzing densely packed groups of nuclei (Figure 23d), offering insights into their collective behaviors and potential biological interactions.

The network analysis yields several key visual outputs and quantitative data, like degree of connectivity, figure 24a reveals the degree of connectivity among the




(a) Original Image

(b) Waxman Graph

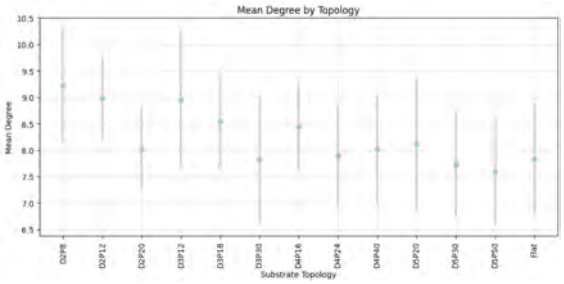
(c) DBSCAN Analysis

(d) Degree of Connection

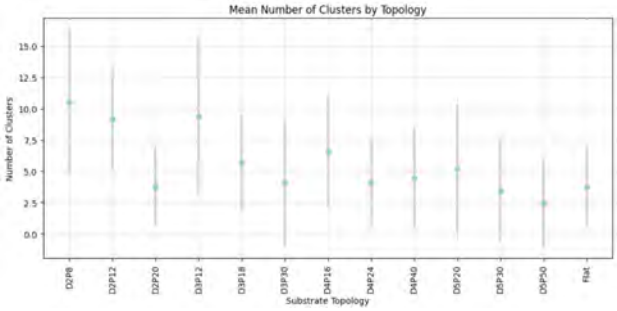
Figure 23: Network Analysis for nuclei sample on substrate topology of pillars of diameter 2 and pitch 20

nuclei, providing a measure of how interconnected each nucleus is within the overall network. This metric is vital for understanding the robustness and vulnerability of the cellular network.

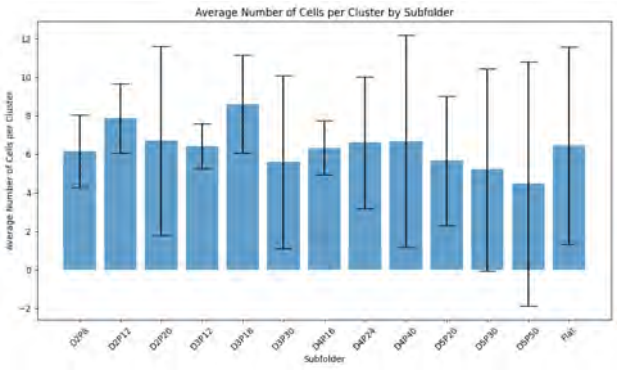
Additionally, figure 24b and 24c provide data on the number and size of clusters identified by DBSCAN, crucial for assessing the aggregation tendency of nuclei under different conditions.




(a) Mean Degree of Connection per Substrate Topology




(b) Number of clusters Identified by DBSCAN




(c) Number of Cells per Cluster Identified by DBSCAN

Figure 24: Data Analysis of Graph obtained from the different topologies

## 7. Discussion

### *7.1. Milestone 1.1: Nuclei and cluster identification*

The successful deployment of traditional image analysis techniques in our study has led to highly accurate segmentation results, crucial for the training of the Mask-RCNN model. As illustrated in Figure 11, our methodology effectively identifies clusters and single nuclei on substrates characterized by very fine spatial features such as pillars, underscoring the importance of high-resolution differentiation in studying cellular responses to micro-environmental features. This mirrors insights from studies like those by de Santos-Sierra et al. (2014), which explored how small-world network configurations emerge in in vitro neuronal cultures, reflecting efficient neuronal processing and network resilience.

The analysis, as depicted in Figure 12, reveals how varying substrate topographies influence cluster formation, with smaller configurations like a diameter of 2 $\mu$m and pitch of 8 $\mu$m (D2P8) showing a higher propensity for cluster formation compared to larger or flat topographies. This suggests that tighter substrate spacing may enhance cell-to-cell interactions or constrain space, encouraging closer cellular aggregation. Such findings are pivotal as they demonstrate how micro-environmental conditions can mimic the dense cellular environments found in natural tissues, potentially affecting cellular behavior and interactions as neurons typically cluster in specific patterns crucial for brain function and organization.

Further, the detailed image analysis in Figure 11 showcases the ability to pinpoint both individual nuclei and larger clusters using advanced processing techniques, highlighting structures that conventional tools like ImageJ (Abràmoff et al., 2004) or StarDist2D Schmidt et al. (2018) might miss. This capability is critical for understanding the complex cellular arrangements that can occur on micro-patterned substrates, where precise segmentation is essential.

The training process of our Mask-RCNN model, indicated by the declining loss metrics in Figure 13 and the high IoU scores—0.901 for boxes and 0.852 for masks—emphasizes the model's accuracy in detecting and segmenting nuclei, crucial for identifying clusters accurately as shown in Figure 14.

By integrating and extending sophisticated image analysis techniques, our research not only aligns with but also builds upon foundational studies like that of de Santos-Sierra et al. (2014), enhancing our understanding of how substrate topographies and micro-environmental cues influence neuronal network dynamics. This approach underscores the potential of these methodologies to be adapted across various neuronal culture conditions and substrate types, broadening their application in neural tissue engineering and related scientific fields.

### *7.2. Milestone 1.2: Soma and neurite identification*

The segmentation method developed for this project is specifically designed for images with fluorescent pillar substrates, characterized by inherently noisy backgrounds. This method utilizes a detailed preprocessing workflow that includes denoising and contrast enhancement, followed by a dual-segmentation process using Watershed and GrabCut algorithms. This ensures precise identification of neuronal structures. Additionally, the segmentation accuracy is further refined through the training of a U-Net model, optimizing the method for detailed analyses within challenging imaging environments.

In contrast, the ANDA tool, designed by Wæhler et al. (2023), focuses on automation and efficiency across various cell types and employs global thresholding, Watershed segmentation, and optional Weka segmentation for low-contrast images, using Fiji's customization capabilities. Although highly adaptable, ANDA may require additional configuration to effectively address the noise levels typical in fluorescent pillar substrates.

To enhance the segmentation of soma and neurites, various U-Net models with different backbone architectures were evaluated. RegNetY320 (Radosavovic et al., 2020) emerged as the preferred model, recording the highest F1-score and IoU, indicating its exceptional capability in accurately delineating complex neuronal structures. While models like VGG16, MobileNET, EfficientNet, and ResNet 132 showed commendable performances, they did not match the effectiveness of RegNetY. The selection of RegNetY highlights the impact of network architecture on segmentation outcomes, particularly in complex imaging scenarios.

Furthermore, the study by Mari et al. (2015) provides a comprehensive framework for the morphological analysis of neurons cultured on microelectrode arrays (MEAs), offering valuable insights for our segmentation approach. Our method extends these analytical techniques to measure normalized neurite lengths and numbers, these metrics are indispensable for understanding neuronal health and network capabilities, offering insights into how different environments affect neuronal morphology and function. This ensures that results are comparable across different samples and conditions. The post-processing techniques employed refine the accuracy of segmentation, as illustrated in the provided figures 17a.

The analysis of neurite metrics across different substrate topologies (Figure 17b)—Mushroom, Stubby, and Thin—reveals how substrate characteristics influence neurite growth. Mushroom substrates show consistent neurite growth with moderate length and low variation. Stubby substrates, in contrast, display significant variation, suggesting they may support extended neurite growth under certain conditions. Thin substrates exhibit the shortest neurite lengths, indicating potential constraints on growth due to their uniform environment.

The study not only reaffirms but also builds upon foundational work, enhancing our understanding of neuronal network dynamics influenced by varying substrate topographies. This demonstrates the potential for applying these advanced image analysis techniques across different neuronal culture conditions, expanding their applicability in neural tissue engineering and related fields.

### *7.3. Milestone 1.3: Neurite tracking*

In the current study, tracing neurite coordinates is effectively accomplished using NeuronJ, where tracing data is converted into continuous coordinate paths. This method ensures that each neurite is represented as a continuous entity, facilitating the precise analysis of neurite paths and their dynamic changes over time. Unlike this approach, the NeuriTES (Mencattini et al., 2021) platform is based on manual labeling in initial frames for network training, which, while effective for static or slow-changing conditions typical in motor neuron studies, may not provide the flexibility required to capture the rapid and unpredictable growth patterns of developing cortical neurons.

DBSCAN (Ester et al., 1996) clustering is applied within each frame to effectively differentiate individual neurites based on the density of traced points. This intra-frame clustering allows for the accurate identification of neurite entities, which is crucial for monitoring the morphological changes typical of rapidly developing neurons. In comparison, the NeuriTES platform, which does not inherently focus on density-based clustering, may struggle to differentiate closely packed or rapidly evolving neurite structures. The adaptability of DBSCAN to changes in neurite density and arrangement offers significant advantages in tracking cortical neurons, which exhibit high variability and faster dynamics during development.

Further, inter-frame DBSCAN clustering is utilized to track the persistence and evolution of neurites over time, an approach that is particularly useful for observing developmental changes and interactions. This method contrasts with NeuriTES, which might not adequately address the high variability and rapid dynamics of cortical neurons, whose growth patterns can significantly alter between imaging sessions.

Our DBSCAN-based approach, inspired by the method described by Kim and Cho (2021) in their study on multi-object tracking, applies this robust clustering technique effectively in neuron tracking. Their research highlighted DBSCAN's efficacy in enhancing multi-object tracking by reducing noise vulnerability and simplifying the data association process, which we adapted to suit the complex and dynamic environment of cortical neuron tracking in live imaging.

The analysis of the neurite tracking results, demonstrated in various figures, sheds light on neurite behavior in terms of growth dynamics. Length changes, directional movements, and growth velocities of neurites are quantitatively tracked, as depicted in the figures showing initial states and subsequent transformations over time (Figure 18).

Furthermore, the plots in figure 19 delineate the average length changes, direction and growth velocities of neurites over time, contrasting neurites in contact with pillars to those in flat areas. It shows that neurites in contact with pillars generally exhibit more significant length fluctuations compared to those in flat areas (Figure 19a). This suggests that the micro-environmental features of the pillars might either promote or inhibit

neurite elongation depending on the local conditions and interactions at the cellular level. Neurites in contact with pillars demonstrate a pattern of sharper peaks and deeper troughs in length change, indicating more dynamic growth behavior.

The average neurite growth velocity further supports this observation. Neurites in contact with pillars show higher velocity fluctuations (Figure 19c), underscoring a more active response to the textured environment provided by the pillars. This can be interpreted as neurites rapidly adapting their growth strategies in response to the physical cues presented by the pillar structures.

Additionally, the plot 19b, red lines, representing neurites in contact with pillars, predominantly show more extended vectors indicating both longer movements and more varied directional changes. This suggests that neurites interacting with pillars exhibit not only more dynamic movement but also greater exploratory behavior, potentially adapting to the micro-topographical cues provided by the pillars. These neurites display a broader spread of angles, which could imply a more complex environment where neurites continuously adjust their growth paths in response to physical contacts and spatial constraints imposed by the pillars. Conversely, the blue vectors representing neurites in flat areas are shorter and more concentrated around specific angles. This pattern indicates that neurites in flat areas may experience less physical interaction with their environment, leading to more linear and predictable growth patterns. The lack of substantial directional changes could suggest a more uniform and less challenging environment, where neurites can extend without the need to navigate around physical obstacles.

Despite the robust methodology employed, the Mask-RCNN model's segmentation performance, indicated by average IoU scores, suggests room for improvement. The moderate accuracy in segmentation points to potential enhancements in model training or data quality, which could further refine the understanding of neurite dynamics.

The methodologies employed in this study are well-suited for tracking the dynamic and unpredictable growth patterns of cortical neurons in developmental stages, offering significant advantages in terms of flexibility and adaptability over the NeuriTES approach. While NeuriTES provides robust tools for studying motor neuron degeneration, its techniques may not fully capture the rapid and variable changes characteristic of developing cortical neurons.

### *7.4. Milestone 2: Protein expression*

In the protein expression analysis methodology, images are systematically processed to extract individual fluorescence channels, specifically targeting regions of interest such as soma, neurites, and their combined areas. This segmentation is facilitated by applying masks to the original images, allowing for precise isolation of these regions. The calculation of mean intensities for both cells and background within these masks enables a more accurate measurement of protein expression levels.

In this study, the protein expression analysis aimed to elucidate the influence of substrate topographies on the localization and concentration of key adhesion proteins, Integrin and Paxillin. These proteins are crucial for forming focal adhesions, which play significant roles in neurite outgrowth, axon guidance, and neuronal migration during brain development. Paxillin, in particular, integrates integrin and growth factor signaling pathways that coordinate the cytoskeletal rearrangements required for neurite initiation and extension (Chang et al., 2017).

The heatmap for Integrin expression (Figure 22) shows notable variation in fluorescence intensity across different substrate topologies. Integrin exhibits the highest mean fluorescence intensity on mushroom-type pillars with a pitch of 30 (P30), indicating strong cellular adhesion and active engagement with these substrate features. This suggests that mushroom-type pillars with larger pitches provide an optimal environment for integrin-mediated adhesion.

Substrates with stubby and thin pillars show lower fluorescence intensities for Integrin, implying less active adhesion processes. Flat surfaces, serving as a control, also display relatively low Integrin expression. These observations highlight the role of substrate topography in modulating cell-substrate interactions, where more complex features like mushroom-type pillars significantly enhance adhesion protein activity.

Similarly, the heatmap for Paxillin expression (Figure 21) demonstrates a comparable pattern to Integrin. Paxillin shows the highest fluorescence intensity on mushroom-type pillars with a pitch of 30 (P30), underscoring that these substrates promote robust focal adhesion sites. This intense Paxillin expression supports the notion that mushroom-type pillars create a conducive environment for cell adhesion.

Substrates with stubby and thin pillars display moderate Paxillin expression, whereas flat surfaces exhibit the lowest levels. The correlation between substrate complexity and Paxillin intensity suggests that more intricate topographies provide better support for the formation and maintenance of focal adhesions, which are critical for cellular stability and signaling.

This analysis revealed that mushroom-type pillars with larger pitches (P30) promote higher levels of both Integrin and Paxillin, enhancing cellular adhesion. These findings align with the known effects of substrate stiffness on protein expression, where softer substrates promote higher Paxillin expression and endocytosis, while stiffer substrates favor integrin localization to focal adhesions (Verma et al., 2021). By understanding these interactions and leveraging advanced image analysis tools, substrates can be designed to optimize cell behavior for specific applications, enhancing neural

tissue engineering and other biomedical fields.

### *7.5. Milestone 3: Network analysis of nuclei*

The network analysis of nuclei was conducted using the StarDist2D model Schmidt et al. (2018); Weigert and Schmidt (2022), which is designed for identifying single nuclei in biological images (Figure 23a). Following segmentation, a Waxman graph was created using the centroids of the segmented nuclei (Figure 23b), which facilitated the study of the degree of connection of the identified structures (Figure 23d). Cluster analysis was performed using DBSCAN (Density-Based Spatial Clustering of Applications with Noise), analyzing the spatial distributions and cluster formations among the nuclei (Figure 23c).

The results of the network analysis are depicted in several figures. The mean degree of connectivity (Figure 24a) indicates that the smallest topography (D2P8) exhibits a higher degree of connection compared to larger or flat surfaces. This aligns with the cluster formation analysis, where smaller topographies tend to promote more intense clustering.

Regarding the average number of cells per cluster (Figure 24c), the data shows that while the clusters formed on small topographies consist of fewer cells, they are more consistent. This is reflected in the lower variability observed in these substrates compared to larger or flat surfaces. Additionally, the mean number of clusters identified by DBSCAN (Figure 24b) shows a higher number of clusters in smaller topographies, confirming the tendency for more frequent clustering in these conditions.

It is important to note that the DBSCAN analysis was specifically focused on clusters made up of single nuclei, as identified by StarDist2D. This model is highly effective for single nuclei identification but less so for dense clusters of nuclei, highlighting the necessity of complementary methods for comprehensive analysis.

The approach was inspired by the research of Onesto et al. (2019), which investigated the formation of cortical-like mini-columns on zinc oxide nanowire surfaces and their sensitivity to topographical features. Their findings emphasized the importance of substrate topography in guiding neuronal cell assembly into clusters with high connectivity and small-world network attributes. This research provided a foundational understanding that informed the use of the Waxman graph and DBSCAN for analyzing the spatial organization of nuclei in the current experiments.

In summary, the network analysis underscored the critical role of substrate topographies in influencing the spatial organization and connectivity of neuronal networks. By employing advanced segmentation models and graph-based analysis techniques, the study provides a comprehensive framework for understanding the complex dynamics of neuronal network formation and their dependence on the physical environment. Drawing inspiration from previous research, this integrated approach highlights the significance of merging computational methods with biological insights to enhance the knowledge of neuronal network behavior.

## 8. Conclusions

### *8.1. Milestone 1.1: Nuclei and cluster identification*

By leveraging traditional image analysis tools, highly accurate segmentations, a prerequisite for the training of the Mask-RCNN model, were achieved. It was demonstrated that this approach performs well, even on substrates containing fine spatial features, which are critical in studies monitoring cell responses. These results indicate that micro-environmental conditions, particularly at high densities with tight substrate configurations, enhance cell-cell interactions and clustering. Furthermore, the complex cellular arrangements, which are not resolvable by most common tools, were captured by the advanced processing techniques used in the current approach, thus validating its capability for studying complex neuronal networks.

### *8.2. Milestone 1.2: Soma and neurite identification*

For images with fluorescent pillar substrates, a segmentation protocol was developed and applied successfully even in highly noisy contexts. This was achieved through the implementation of a thorough preprocessing algorithm and double segmentation. The use of a superior model, RegNetY, facilitated an enhancement in segmentation accuracy. By extending beyond standard analysis procedures to calculate normalized neurite lengths and numbers, and by interpreting changes in neurite outgrowth sensitivity to substrate topographies, the applicability of this approach under different conditions in neuronal cultures was validated.

### *8.3. Milestone 1.3: Neurite tracking*

The use of NeuronJ and DBSCAN clustering in the methodology for neurite tracking enabled the dynamics of neurites over time to be tracked, offering significant improvements over other methods in capturing rapid and unpredictable growth patterns during cortical neuron development. Analyses have shown that neurites in contact with pillars exhibit increased fluctuations in length and velocity, responding dynamically to textured environments. This methodology is adaptable for studying neuronal development and underscores the benefits of advanced clustering techniques in neuron tracking.

### *8.4. Milestone 2: Protein expression*

In the protein expression assay, it was observed that mushroom-type pillars with larger pitches promoted higher levels of Integrin and Paxillin, enhancing cellular adhesion. Extrapolating from the effects of substrate

stiffness on protein expression, it is expected that more complex topographies better support focal adhesion formation. The role of substrate design in achieving optimal cell behavior for custom applications was claimed by this work, which integrates traditional imaging with advanced computational methods to provide a comprehensive view of cellular dynamics.

### *8.5. Milestone 3: Network analysis of nuclei*

The network analysis of nuclei described the impact of different substrate topographies on changes in the spatial structure and connectivity of neuronal networks. Smaller topographies resulted in an increased rate of connectivity and more consistent clustering. The framework provided a comprehensive understanding of neuronal network dynamics through advanced segmentation models combined with graph-based analysis techniques. Based on earlier studies, this approach substantiates the need to integrate computational methodologies with biological insights to advance the knowledge of neuronal network behavior.

Each milestone underscores the effectiveness of advanced image analysis and computational techniques in elucidating cell-neuronal network interactions, highlighting how physical environments can influence neuronal behavior, with significant implications in neural tissue engineering and related fields.

## References


Abràmoff, M.D., Magalhães, P.J., Ram, S.J., 2004. Image processing with imagej. Biophotonics international 11, 36–42.

Abu-Ain, W., Abdullah, S.N.H.S., Bataineh, B., Abu-Ain, T., Omar, K., 2013. Skeletonization algorithm for binary images. Procedia Technology 11, 704–709.

Adewole, D.O., Serruya, M.D., Wolf, J.A., Cullen, D.K., 2019. Bioactive neuroelectronic interfaces. Frontiers in neuroscience 13, 442841.

Al-Kofahi, O., Radke, R.J., Roysam, B., Banker, G., 2006. Automated semantic analysis of changes in image sequences of neurons in culture. IEEE Transactions on Biomedical Engineering 53, 1109–1123.

Aspiotis, V., Miltiadous, A., Kalafatakis, K., Tzimourta, K.D., Giannakeas, N., Tsipouras, M.G., Peschos, D., Glavas, E., Tzallas, A.T., 2022. Assessing electroencephalography as a stress indicator: A vr high-altitude scenario monitored through eeg and ecg. Sensors 22, 5792.

Banker, G., 2018. The development of neuronal polarity: a retrospective view. Journal of Neuroscience 38, 1867–1873.

Batool, S., Raza, H., Zaidi, J., Riaz, S., Hasan, S., Syed, N.I., 2019. Synapse formation: from cellular and molecular mechanisms to neurodevelopmental and neurodegenerative disorders. Journal of neurophysiology .

Bellingacci, L., Mancini, A., Gaetani, L., Tozzi, A., Parnetti, L., Di Filippo, M., 2021. Synaptic dysfunction in multiple sclerosis: a red thread from inflammation to network disconnection. International Journal of Molecular Sciences 22, 9753.

Bokel, C., Brown, N.H., 2002. Integrins in development: moving on, responding to, and sticking to the extracellular matrix. Dev Cell 3, 311–321.

Chang, T.Y., Chen, C., Lee, M., Chang, Y.C., Lu, C.H., Lu, S.T., Wang, D.Y., Wang, A., Guo, C.L., Cheng, P.L., 2017. Paxillin facilitates timely neurite initiation on soft-substrate environments by interacting with the endocytic machinery. Elife 6, e31101.

Cho, Y.H., Park, Y.G., Kim, S., Park, J.U., 2021. 3d electrodes for bioelectronics. Advanced Materials 33, 2005805.

Claverol-Tinture, E., Ghirardi, M., Fiumara, F., Rosell, X., Cabestany, J., 2005. Multielectrode arrays with elastomeric microstructured overlays for extracellular recordings from patterned neurons. Journal of neural engineering 2, L1.

Cuttaz, E.A., Bailey, Z.K., Chapman, C.A., Goding, J.A., Green, R.A., 2024. Polymer bioelectronics: A solution for both stimulating and recording electrodes. Advanced Healthcare Materials , 2304447.

Delgado-García, J.M., 2015. Cajal and the conceptual weakness of neural sciences. Frontiers in Neuroanatomy 9, 128.

Drakopoulou, S., Varkevisser, F., Sohail, L., Aqamolaei, M., Costa, T.L., Spyropoulos, G.D., 2023. Hybrid neuroelectronics: towards a solution-centric way of thinking about complex problems in neurostimulation tools. frontiers in electronics 4, 1250655.

Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise, in: kdd, pp. 226–231.

Fanelli, A., Ferlauto, L., Zollinger, E.G., Brina, O., Reymond, P., Machi, P., Ghezzi, D., 2022. Transient neurovascular interface for minimally invasive neural recording and stimulation. Advanced Materials Technologies 7, 2100176.

Friedrich, R.W., Genoud, C., Wanner, A.A., 2013. Analyzing the structure and function of neuronal circuits in zebrafish. Frontiers in neural circuits 7, 71.

Fujita, H., Oikawa, R., Hayakawa, M., Tomoike, F., Kimura, Y., Okuno, H., Hatashita, Y., Oliveros, C.F., Bito, H., Ohshima, T., et al., 2020. Quantification of native mrna dynamics in living neurons using fluorescence correlation spectroscopy and reduction-triggered fluorescent probes. Journal of Biological Chemistry 295, 7923–7940.

Ganguli, M.P., Upton, A.R., Kamath, M., 2017. Deep brain stimulation as a treatment for refractory epilepsy: Review of the current state-of-the-art. Journal of Long-Term Effects of Medical Implants 27.

Gärtner, A., Fornasiero, E.F., Dotti, C.G., 2015. Cadherins as regulators of neuronal polarity. Cell adhesion & migration 9, 175–182.

Go, G.T., Lee, Y., Seo, D.G., Lee, T.W., 2022. Organic neuroelectronics: from neural interfaces to neuroprosthetics. Advanced Materials 34, 2201864.

Gu, X., Jia, C., Wang, J., 2023. Advances in understanding the molecular mechanisms of neuronal polarity. Molecular Neurobiology 60, 2851–2870.

Guimerà-Brunet, A., Masvidal-Codina, E., Cisneros-Fernández, J., Serra-Graells, F., Garrido, J.A., 2021. Novel transducers for high-channel-count neuroelectronic recording interfaces. Current opinion in biotechnology 72, 39–47.

und Halbach, O.v.B., 2009. Structure and function of dendritic spines within the hippocampus. Annals of Anatomy-Anatomischer Anzeiger 191, 518–531.

Hales, C.M., Rolston, J.D., Potter, S.M., 2010. How to culture, record and stimulate neuronal networks on micro-electrode arrays (meas). JoVE (Journal of Visualized Experiments) , e2056.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, pp. 2961–2969.

Higgs, V.E., Das, R.M., 2022. Establishing neuronal polarity: microtubule regulation during neurite initiation. Oxford Open Neuroscience 1, kvac007.

Holtzheimer, P.E., Husain, M.M., Lisanby, S.H., Taylor, S.F., Whitworth, L.A., McClintock, S., Slavin, K.V., Berman, J., McKhann, G.M., Patil, P.G., et al., 2017. Subcallosal cingulate deep brain stimulation for treatment-resistant depression: a multisite, randomised, sham-controlled trial. The Lancet Psychiatry 4, 839–849.

Hooi, L.L., Fitzrol, D.N., Rajapathy, S.K., Chin, T.Y., Halim, S.A., Kandasamy, R., Hassan, W.M.N.W., Idris, B., Ghani, A.R.I., Idris, Z., et al., 2017. Deep brain stimulation (dbs) for movement disorders: an experience in hospital universiti sains malaysia (husm) involving 12 patients. The Malaysian journal of medical sciences: MJMS 24, 87.

Iakubovskii, P., 2019. Segmentation models pytorch.

Illingworth, J., Kittler, J., 1987. The adaptive hough transform. IEEE Transactions on Pattern Analysis and Machine Intelligence , 690–698.

Kandel, E.R., Schwartz, J.H., Jessell, T.M., Siegelbaum, S., Hudspeth, A.J., Mack, S., et al., 2000. Principles of neural science. volume 4. McGraw-hill New York.

Kaur, H., Siwal, S.S., Saini, R.V., Singh, N., Thakur, V.K., 2022. Significance of an electrochemical sensor and nanocomposites: Toward the electrocatalytic detection of neurotransmitters and their importance within the physiological system. ACS Nanoscience Au 3, 1–27.

Keogh, C., 2020. Optimizing the neuron-electrode interface for chronic bioelectronic interfacing. Neurosurgical Focus 49, E7.

Khan, S.P., Auner, G.G., Palyvoda, O., Newaz, G.M., 2011. Biocompatibility assessment of next generation materials for brain implantable microelectrodes. Materials letters 65, 876–879.

Kim, J., Cho, J., 2021. Dbscan-based tracklet association annealer for advanced multi-object tracking. Sensors 21, 5715.

Kim, K.M., Son, K., Palmore, G.T.R., 2015. Neuron image analyzer: Automated and accurate extraction of neuronal data from low quality images. Scientific Reports 5, 17062. doi:10.1038/srep17062.

Kireev, D., Offenhäusser, A., 2018. Graphene & two-dimensional devices for bioelectronics and neuroprosthetics. 2D Materials 5, 042004.

Kireev, D., Sel, K., Ibrahim, B., Kumar, N., Akbari, A., Jafari, R., Akinwande, D., 2022. Continuous cuffless monitoring of arterial blood pressure via graphene bioimpedance tattoos. Nature nanotechnology 17, 864–870.

Krook-Magnuson, E., Gelinas, J.N., Soltesz, I., Buzsáki, G., 2015. Neuroelectronics and biooptics: closed-loop technologies in neurological disorders. JAMA neurology 72, 823–829.

Kumar, S., Yadav, S., Kumar, A., 2024. Accuracy of oscillometric-based blood pressure monitoring devices: impact of pulse volume, arrhythmia, and respiratory artifact. Journal of Human Hypertension 38, 45–51.

Lamprecht, M.R., Sabatini, D.M., Carpenter, A.E., 2007. Cellprofiler™: free, versatile software for automated biological image analysis. Biotechniques 42, 71–75.

Leshchyns' Ka, I., Sytnyk, V., 2016. Reciprocal interactions between cell adhesion molecules of the immunoglobulin superfamily and the cytoskeleton in neurons. Frontiers in cell and developmental biology 4, 9.

Li, Q., Wang, L., Ma, Y., Yue, W., Zhang, D., Li, J., 2019. P-rex1 overexpression results in aberrant neuronal polarity and psychosis-related behaviors. Neuroscience Bulletin 35, 1011–1023.

Liang, Y., Offenhäusser, A., Ingebrandt, S., Mayer, D., 2021. Pedot: Pss-based bioelectronic devices for recording and modulation of electrophysiological and biochemical cell signals. Advanced Healthcare Materials 10, 2100061.

Liu, Z., Jin, L., Chen, J., Fang, Q., Ablameyko, S., Yin, Z., Xu, Y., 2021. A survey on applications of deep learning in microscopy image analysis. Computers in biology and medicine 134, 104523.

López-Colomé, A.M., Lee-Rivera, I., Benavides-Hidalgo, R., López, E., 2017. Paxillin: a crossroad in pathological cell migration. Journal of hematology & oncology 10, 1–15.

Luan, L., Yin, R., Zhu, H., Xie, C., 2023. Emerging penetrating neural electrodes: In pursuit of large scale and longevity. Annual Review of Biomedical Engineering 25, 185–205.

Mari, J.F., Saito, J.H., Neves, A.F., Lotufo, C.M.d.C., Destro-Filho, J.B., Nicoletti, M.d.C., 2015. Quantitative analysis of rat dorsal root ganglion neurons cultured on microelectrode arrays based on fluorescence microscopy image processing. International journal of neural systems 25, 1550033.

Mariano, A., Lubrano, C., Bruno, U., Ausilio, C., Dinger, N.B., Santoro, F., 2021. Advances in cell-conductive polymer biointerfaces and role of the plasma membrane. Chemical reviews 122, 4552–4580.

Martínez, G., Howard, N., Abbott, D., Lim, K., Ward, R., Elgendi, M., 2018. Can photoplethysmography replace arterial blood pressure in the assessment of blood pressure? Journal of clinical medicine 7, 316.

Matino, L., Rastogi, S.K., Garma, L.D., Cohen-Karni, T., Santoro, F., 2020. Characterization of the coupling between out-of-plane graphene and electrogenic cells. Advanced Materials Interfaces 7, 2000699.

Mencattini, A., Spalloni, A., Casti, P., Comes, M.C., Di Giuseppe, D., Antonelli, G., d'Orazio, M., Filippi, J., Corsi, F., Isambert, H., et al., 2021. Neurites. monitoring neurite changes through transfer entropy and semantic segmentation in bright-field time-lapse microscopy. Patterns 2.

Milos, F., Belu, A., Mayer, D., Maybeck, V., Offenhäusser, A., 2021. Polymer nanopillars induce increased paxillin adhesion assembly and promote axon growth in primary cortical neurons. Advanced biology 5, 2000248.

Muthmann, J.O., Amin, H., Sernagor, E., Maccione, A., Panas, D., Berdondini, L., Bhalla, U.S., Hennig, M.H., 2015. Spike detection for large neural populations using high density multielectrode arrays. Frontiers in neuroinformatics 9, 28.

Nella, K.T., Norton, B.M., Chang, H.T., Heuer, R.A., Roque, C.B., Matsuoka, A.J., 2022. Bridging the electrode–neuron gap: finite element modeling of in vitro neurotrophin gradients to optimize neuroelectronic interfaces in the inner ear. Acta Biomaterialia 151, 360–378.

Neumann, W.J., Steiner, L.A., Milosevic, L., 2023. Neurophysiological mechanisms of deep brain stimulation across spatiotemporal resolutions. Brain 146, 4456–4468.

Nosov, G., Kahms, M., Klingauf, J., 2020. The decade of super-resolution microscopy of the presynapse. Frontiers in synaptic neuroscience 12, 32.

Onesto, V., Villani, M., Narducci, R., Malara, N., Imbrogno, A., Allione, M., Costa, N., Coppedè, N., Zappettini, A., Cannistraci, C., et al., 2019. Cortical-like mini-columns of neuronal cells on zinc oxide nanowire surfaces. Scientific reports 9, 4021.

Opstad, I.S., Ströhl, F., Fantham, M., Hockings, C., Vanderpoorten, O., van Tartwijk, F.W., Lin, J.Q., Tinguely, J.C., Dullo, F.T., Kaminski-Schierle, G.S., et al., 2020. A waveguide imaging platform for live-cell tirf imaging of neurons over large fields of view. Journal of Biophotonics 13, e201960222.

O'shea, K., Nash, R., 2015. An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458 .

Ossinger, A., Bajic, A., Pan, S., Andersson, B., Ranefall, P., Hailer, N.P., Schizas, N., 2020. A rapid and accurate method to quantify neurite outgrowth from cell and tissue cultures: Two image analytic approaches using adaptive thresholds or machine learning. Journal of Neuroscience Methods 331, 108522.

Parekh, P.K., Johnson, S.B., Liston, C., 2022. Synaptic mechanisms regulating mood state transitions in depression. Annual Review of Neuroscience 45, 581–601.

Paternò, G.M., Bondelli, G., Lanzani, G., 2021. Bringing microbiology to light: toward all-optical electrophysiology in bacteria. Bioelectricity 3, 136–142.

Pitsis, G., 2018. Design and Implementation of an FPGA-Based Convolutional Neural Network Accelerator. Ph.D. thesis.

Qi, Y., Kang, S.K., Fang, H., Editors, G., 2023. Advanced materials for implantable neuroelectronics. MRS bulletin 48, 475–483.

Radosavovic, I., Kosaraju, R.P., Girshick, R., He, K., Dollár, P., 2020. Designing network design spaces.

Radotić, V., Braeken, D., Kovačić, D., 2017. Microelectrode array-induced neuronal alignment directs neurite outgrowth: analysis using a fast fourier transform (fft). European biophysics journal 46, 719–727.

Rafiq, N.B.M., Lyons, L.L., Gowrishankar, S., Camilli, P.D., Ferguson, S.M., 2022. Jip3 links lysosome transport to regulation of multiple components of the axonal cytoskeleton. Communications Biology 5. URL: https://api.semanticscholar.org/CorpusID:257113458.

Reddy, S., Xiao, Q., Liu, H., Li, C., Chen, S., Wang, C., Chiu, K., Chen, N., Tu, Y., Ramakrishna, S., et al., 2019. Bionanotube/poly (3, 4-ethylenedioxythiophene) nanohybrid as an electrode for the neural interface and dopamine sensor. ACS applied materials & interfaces 11, 18254–18267.

Reza, A.M., 2004. Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. Journal of VLSI signal processing systems for signal, image and video technology 38, 35–44.

Rinklin, P., Wolfrum, B., 2021. Recent developments and future perspectives on neuroelectronic devices. Neuroforum 27, 213–224.

Rossi, L.F., Kullmann, D.M., Wykes, R.C., 2018. The enlightened brain: novel imaging methods focus on epileptic networks at multiple scales. Frontiers in cellular neuroscience 12, 82.

Roy, A.G., Navab, N., Wachinger, C., 2018. Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks. IEEE transactions on medical imaging 38, 540–549.

Rudzki, L., Maes, M., 2021. From "leaky gut" to impaired glia-neuron communication in depression, in: Major Depressive Disorder: Rethinking and Understanding Recent Discoveries. Springer, pp. 129–155.

de Santos-Sierra, D., Sendina-Nadal, I., Leyva, I., Almendral, J.A., Anava, S., Ayali, A., Papo, D., Boccaletti, S., 2014. Emergence of small-world anatomical networks in self-organizing clustered neuronal cultures. PloS one 9, e85828.

Schiavone, G., Kang, X., Fallegger, F., Gandar, J., Courtine, G., Lacour, S.P., 2020. Guidelines to study and develop soft electrode systems for neural stimulation. Neuron 108, 238–258.

Schmidt, U., Weigert, M., Broaddus, C., Myers, G., 2018. Cell detection with star-convex polygons, in: Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II, pp. 265–273.

Schmued, L., Kyriakidis, K., Fallon, J., Ribak, C., 1989. Neurons containing retrogradely transported fluoro-gold exhibit a variety of lysosomal profiles: a combined brightfield, fluorescence, and electron microscopic study. Journal of neurocytology 18, 333–343.

Schurr, J., Haghofer, A., Lanzerstorfer, P., Winkler, S., 2023. Automated segmentation of patterned cells in micropatterning microscopy images, in: Roque, A.C.A., Gracanin, D., Lorenz, R., Tsanas, A., Bier, N., Fred, A., Gamboa, H. (Eds.), Biomedical Engineering Systems and Technologies, Springer Nature Switzerland, Cham. p. 34–52.

Seo, K.J., Hill, M., Ryu, J., Chiang, C.H., Rachinskiy, I., Qiang, Y., Jang, D., Trumpis, M., Wang, C., Viventi, J., et al., 2023. A soft, high-density neuroelectronic array. Npj flexible electronics 7, 40.

Shan, Y., Farmer, S.M., Wray, S., 2021. Drebrin regulates cytoskeleton dynamics in migrating neurons through interaction with cxcr4. Proceedings of the National Academy of Sciences 118, e2009493118.

Solecki, D.J., 2022. Neuronal polarity pathways as central integrators of cell-extrinsic information during interactions of neural progenitors with germinal niches. Frontiers in Molecular Neuroscience 15, 829666.

Sommer, C., Straehle, C., Koethe, U., Hamprecht, F.A., 2011. Ilastik: Interactive learning and segmentation toolkit, in: 2011 IEEE international symposium on biomedical imaging: From nano to macro, IEEE. pp. 230–233.

Szabo, B., Starke, K., 2021. Synaptic transmission, in: Encyclopedia of Molecular Pharmacology. Springer, pp. 1–9.

Takano, T., Funahashi, Y., Kaibuchi, K., 2019. Neuronal polarity: positive and negative feedback signals. Frontiers in cell and developmental biology 7, 69.

Tang-Schomer, M.D., Hu, X., Hronik-Tupaj, M., Tien, L.W., Whalen, M.J., Omenetto, F.G., Kaplan, D.L., 2014. Film-based implants for supporting neuron–electrode integrated interfaces for the brain. Advanced functional materials 24, 1938–1948.

Temiz, Y., Ferretti, A., Leblebici, Y., Guiducci, C., 2012. A comparative study on fabrication techniques for on-chip microelectrodes. Lab on a Chip 12, 4920–4928.

Tong, L., Langton, R., Glykys, J., Baek, S., 2021. Anmaf: an automated neuronal morphology analysis framework using convolutional neural networks. Scientific reports 11, 8179.

Verma, B.K., Chatterjee, A., Kondaiah, P., Gundiah, N., 2021. Substrate stiffness modulates integrin $\alpha 5$ expression and ecm-associated gene expression in fibroblasts. bioRxiv , 2021–11.

Vincent, L., Soille, P., 1991. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Transactions on Pattern Analysis & Machine Intelligence 13, 583–598.

Viswam, V., Obien, M.E.J., Franke, F., Frey, U., Hierlemann, A., 2019. Optimal electrode size for multi-scale extracellular-potential recording from neuronal assemblies. Frontiers in neuroscience 13, 453606.

Wæhler, H.A., Labba, N.A., Paulsen, R.E., Sandve, G.K., Eskeland, R., 2023. Anda: an open-source tool for automated image analysis of in vitro neuronal cells. BMC neuroscience 24, 56.

Wang, Z., Lv, Y., Wu, R., Zhang, Y., 2023. Review of grabcut in image processing. Mathematics 11, 1965.

Waxman, B.M., 1988. Routing of multipoint connections. IEEE journal on selected areas in communications 6, 1617–1622.

Weigert, M., Schmidt, U., 2022. Nuclei instance segmentation and classification in histopathology images with stardist, in: The IEEE International Symposium on Biomedical Imaging Challenges (ISBIC). doi:10.1109/ISBIC56247.2022.9854534.

Yin, X.X., Sun, L., Fu, Y., Lu, R., Zhang, Y., 2022. [retracted] u-net-based medical image segmentation. Journal of healthcare engineering 2022, 4189781.

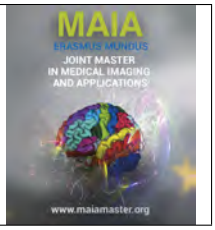
Yousefi, J., 2011. Image binarization using otsu thresholding algorithm. Ontario, Canada: University of Guelph 10.

Zhou, L.Y., Han, F., Qi, S.B., Ma, J.J., Ma, Y.X., Xie, J.L., Zhang, H.C., Fu, X.Y., Chen, J.Q., Li, B., et al., 2020. Inhibition of pten activity promotes ib4-positive sensory neuronal axon growth. Journal of Cellular and Molecular Medicine 24, 11012–11017.

Zhou, Z., Kuo, H.C., Peng, H., Long, F., 2018. Deepneuron: an open deep learning toolbox for neuron tracing. Brain informatics 5, 1–9.

Medical Imaging and Applications

Master Thesis, June 2024




# Cardiac Pathology Classification using multimodal MR images and deep learning techniques

Hsham Ngim, Stéphanie BRICQ

*Université de Bourgogne, ImViA laboratory, Dijon France*

**Abstract**

Cardiovascular diseases (CVDs) are the leading cause of mortality worldwide, necessitating advanced diagnostic techniques for early detection and management. Cardiac Magnetic Resonance (CMR) imaging, including Late Gadolinium Enhancement (LGE) and T1 mapping, has emerged as a pivotal tool for detailed cardiac tissue characterization. Despite its advantages, accurately classifying various cardiac pathologies remains a significant challenge due to the complexity of multimodal data and the limitations of traditional imaging techniques. This study proposes a robust classification framework that integrates multimodal CMR imaging data, specifically LGE, native T1 mapping, and post-contrast T1 mapping with advanced deep learning techniques. By incorporating segmentation masks and employing EfficientNet architectures, the framework aims to enhance the accuracy and reliability of cardiac pathology classification. Additionally, uncertainty quantification methods are utilized to evaluate model confidence, thereby improving the robustness of predictions. Comprehensive experiments were conducted on a dataset comprising 202 patients, categorized into four classes: Cardiomyopathy (CMD), Myocardial Infarction (VIA), Hypertrophic Cardiomyopathy (CMH), and Normal. The results show that the integration of segmentation masks further improves model performance by providing detailed anatomical context. The findings of this research highlight the potential of multimodal CMR imaging combined with deep learning to provide a more accurate, efficient, and comprehensive diagnostic tool for cardiac pathology. This work contributes to the advancement of automated cardiac diagnostics, potentially leading to improved patient outcomes through early and precise disease detection.

**Keywords:** Cardiovascular Diseases, Cardiac Magnetic Resonance Imaging, Deep Learning, T1 Mapping, Late Gadolinium Enhancement, EfficientNet, Segmentation, Uncertainty Quantification

## 1. Introduction

Cardiovascular diseases (CVDs) are widespread among the population and often lead to fatal outcomes. Recent survey statistics indicate that the mortality rate is increasing due to factors such as obesity, high cholesterol, high blood pressure, and tobacco use [Swathy and Saruladha (2022)]. According to the World Health Organization (WHO), cardiovascular diseases (CVDs) are the leading cause of death globally. In 2016, approximately 17.9 million people died from CVDs, accounting for 31% of all deaths worldwide. Of these fatalities, 85% were caused by heart attacks and strokes [World Health Organization (2017)]. Cardiac Magnetic Resonance (CMR) is a specialized form of Magnetic Resonance Imaging (MRI) that provides detailed anatomical and functional information about the heart. Techniques like Late Gadolinium Enhancement (LGE) are valuable for identifying myocardial fibrosis, which is critical for assessing myocardial infarction and other cardiac pathologies. LGE is considered the gold standard for quantifying myocardial infarction, but it has limitations in detecting diffuse fibrosis [Arega et al. (2021)]. To address these limitations, parametric mapping methods such as T1 mapping have been developed. T1 mapping quantifies diffuse myocardial fibrosis and characterizes tissue properties by measuring the longitudinal relaxation time of tissue. This technique provides valuable information about tissue composition and health without the use of contrast agents (native T1 mapping). In contrast, post-contrast T1 mapping involves the administration of gadolinium-based contrast

agents, which enhance the differentiation of healthy and diseased myocardial tissues by altering their T1 relaxation times. The combination of native and post-contrast T1 mapping offers a comprehensive approach to assessing myocardial fibrosis and other pathological changes, enhancing the diagnostic accuracy and prognostic capabilities of CMR [Haaf et al. (2016)]. Despite these advancements, accurately classifying the progression of CVDs remains challenging due to the complex differentiation among various CVD conditions, overlapping symptoms, and variability in disease presentation. Traditional imaging techniques and manual analysis are time-consuming and prone to human error. Consequently, there is a growing need for automated and accurate classification methods that can handle the complexity of multimodal data. Recent advancements in deep learning, particularly convolutional neural networks (CNNs), have shown significant promise in addressing these challenges. Deep learning models can leverage large volumes of heterogeneous data from different imaging modalities, such as LGE and T1 mapping, to provide more comprehensive and accurate diagnostic insights. This approach not only improves diagnostic accuracy but also supports early intervention and enhances patient outcomes by enabling a more detailed understanding of the disease [Swathy and Saruladha (2022)]. The objective of this study is to develop a robust classification framework that integrates multimodal CMR imaging data (LGE, native T1 mapping, and post-contrast T1 mapping) with advanced deep learning techniques. By incorporating segmentation masks and utilizing EfficientNet architectures, this study aims to overcome the limitations of traditional methods and enhance the reliability and accuracy of cardiac pathology classification. Additionally, uncertainty quantification methods will be employed to evaluate model confidence and improve the robustness of predictions. In summary, this research aims to contribute to the field of cardiac pathology classification by leveraging multimodal CMR imaging and deep learning techniques to provide a more accurate, efficient, and comprehensive diagnostic tool. The subsequent sections will detail the state of the art, methodologies, experiments, and results of this study.

## 2. State of the Art

In recent years, the use of multimodal magnetic resonance (MR) imaging combined with deep learning techniques has significantly advanced the field of cardiac pathology classification [Xinga et al. (2024)]. This section reviews the current state of the art, summarizing the latest research findings, methodologies, and identifying gaps that this study aims to address.

Cardiovascular diseases (CVDs) are a leading cause of morbidity and mortality worldwide. Advanced imaging techniques such as Cardiovascular Magnetic Resonance (CMR) have become essential tools for diagnosing and evaluating these conditions [(WHO, 2023)]. CMR techniques, including Late Gadolinium Enhancement (LGE) and T1 mapping, offer detailed insights into myocardial tissue properties, aiding in the detection and quantification of myocardial fibrosis and other pathological changes [Swathy and Saruladha (2022)].

### *2.1. Deep Learning in Cardiac Imaging*

The integration of deep learning with CMR imaging has shown promise in automating and enhancing the diagnostic process. Convolutional Neural Networks (CNNs), in particular, have been widely adopted for their ability to analyze complex imaging data. Recent studies have demonstrated the efficacy of T1 mapping in clinical practice, providing a comprehensive review of its applications in detecting myocardial fibrosis and assessing extracellular volume (ECV) [Arega et al. (2021)].

### *2.2. Multimodal Imaging Approaches*

Multimodal imaging, which combines different CMR techniques, has been explored to improve diagnostic accuracy. The combination of LGE and T1 mapping provides a more detailed characterization of myocardial tissue, enhancing the detection of both focal and diffuse myocardial changes [Swathy and Saruladha (2022)]. This multimodal approach leverages the strengths of each modality, offering a more holistic view of cardiac pathology.

#### *2.2.1. Comparative Analysis of Methodologies*

Recent studies have explored various methodologies to integrate multimodal imaging and deep learning for cardiac pathology classification:

Swathy and Saruladha conducted a comparative study using machine learning and deep learning techniques to classify and predict cardiovascular diseases (CVD). They utilized datasets containing various CMR modalities and applied multiple models including support vector machines (SVM), decision trees, and CNNs. The study concluded that deep learning models, particularly CNNs, provided superior performance in terms of accuracy and robustness [Swathy and Saruladha (2022)].

Arega et al leveraged uncertainty estimates to improve segmentation performance in cardiac MR. They integrated uncertainty quantification within their CNN framework to provide more reliable segmentation results. This methodology involved training the CNNs with dropout layers and using Monte Carlo sampling to estimate uncertainty, which was then used to refine the segmentation outputs [Arega et al. (2021)].

Petersen and Lee explored the fusion of T1 mapping and LGE for improved cardiac diagnosis. Their approach involved preprocessing the T1 mapping and LGE images to align them spatially, followed by feature

extraction using CNNs. The features from both modalities were then fused at different stages of the network to enhance the overall diagnostic accuracy [Petersen and Lee (2020)].

He et al proposed a novel deep learning architecture for cardiac disease classification that integrates multimodal CMR data, including LGE and T1 mapping. Their method involves an attention-based fusion mechanism that selectively emphasizes relevant features from each modality, improving the interpretability and performance of the model [He et al. (2021)].

*J* Liu et al. introduced a hybrid model combining CNNs with recurrent neural networks (RNNs) to capture both spatial and temporal features in cardiac MR images. Their study focused on improving the temporal coherence in volumetric data, which is critical for accurate diagnosis in dynamic cardiac imaging [Liu et al. (2022)].

#### *2.2.2. Comparison with Our Methodology*

Our study builds upon these methodologies by integrating multimodal imaging data (LGE, native T1 mapping, and post-contrast T1 mapping) into a comprehensive classification framework. Key differences and improvements include:

- **Segmentation Integration**: Unlike Swathy and Saruladha, who primarily focused on classification without segmentation, our methodology incorporates segmentation masks to provide additional anatomical context, thereby improving classification accuracy.
- **Uncertainty Quantification**: Similar to Arega et al., we include uncertainty quantification to enhance the reliability of our predictions. However, our approach uses a majority voting mechanism to aggregate predictions from different models, ensuring robust final predictions.
- **Feature Fusion**: While Petersen and Lee focused on fusing features from T1 mapping and LGE, our method extends this fusion to include native and post-contrast T1 mapping as well. This multimodality fusion aims to capture a broader range of myocardial characteristics for more accurate classification.
- **Attention Mechanisms**: Inspired by He et al., our methodology incorporates attention mechanisms to selectively highlight important features from each modality, enhancing model interpretability and performance.
- **Temporal Coherence**: Although our primary focus is on spatial features, our methodology can be extended to incorporate temporal features as explored by Liu et al., providing a comprehensive analysis of dynamic cardiac imaging data.

### *2.3. Challenges and Gaps*

Despite the advancements, several challenges remain in the field of cardiac pathology classification using multimodal MR images. One major issue is the variability in imaging protocols and patient anatomy, which can affect the consistency and accuracy of the models [Swathy and Saruladha (2022)]. Additionally, class imbalance in the datasets poses significant challenges, often leading to biased models that underperform on less represented classes [Johnson et al. (2019)].

Another critical gap is the need for effective integration of segmented anatomical information. While segmentation can provide valuable anatomical context, incorporating this information into deep learning models without increasing complexity or computational cost remains a challenge [Yuemeng Li (2021)].

### *2.4. Current Approaches and Innovations*

Recent innovations have focused on addressing these challenges through various approaches. For instance, the use of EfficientNet architectures has been shown to balance accuracy and computational efficiency effectively [Tan and Le (2019)]. Studies have explored the fusion of native and post-contrast T1 mapping with LGE to enhance feature representation and improve classification performance [Petersen and Lee (2020)]. Moreover, uncertainty quantification methods are being integrated to evaluate model confidence and reliability, thereby improving the robustness of predictions [Maddox and Izmailov (2019)].

### *2.5. Critique and Direction for Future Research*

While these advancements are promising, there is still room for improvement. The variability in imaging protocols and the need for robust data preprocessing techniques highlight the importance of standardizing imaging practices. Additionally, future research should focus on developing methods to handle class imbalance more effectively, perhaps through advanced data augmentation techniques or innovative loss functions that penalize class imbalance [He and Garcia (2019)].

In summary, the state of the art in cardiac pathology classification using multimodal MR images and deep learning techniques has made significant strides, yet several challenges remain. This study aims to address these gaps by leveraging the strengths of multimodal imaging and advanced deep learning architectures, ultimately contributing to the improved diagnosis and management of cardiovascular diseases.

## 3. Material and methods

### *3.1. Dataset*

The dataset employed in this study comprises cardiac magnetic resonance (CMR) imaging data from multiple modalities, specifically Late Gadolinium Enhancement (LGE), native T1 mapping, and post-contrast T1

mapping. This comprehensive dataset includes volumetric data from these modalities for a total of 202 patients. The CMR images were collected from various clinical centers across France. Each image acquisition was performed using a Siemens 1.5T MRI scanner, ensuring consistency in imaging parameters and quality. The slices of native and post-contrast T1 mapping images were realigned based on the center of gravity of the area defined by the manually drawn epicardial contour of the left ventricle. This step was crucial for maintaining anatomical consistency across different slices and modalities.

Each patient of T1 Mapping has three short-axis slices (Apical slices, Mid slices, and Basal slices). In contrast, the number of slices in the LGE modality varies from patient to patient, reflecting differences in clinical protocols and patient anatomy. This variability necessitates careful handling during preprocessing and analysis to ensure robust classification performance. The dataset was categorized into four classes corresponding to different cardiac conditions: CMD (Cardiomyopathy): 71 samples, VIA ( Myocardial Infarction): 70 samples, CMH (Hypertrophic Cardiomyopathy): 30 samples, and Normal: 31 samples.

Additionally, the dataset underwent segmentation as part of a separate project. Manual annotations are provided for each case, including the left ventricular blood pool, myocardium, and right ventricular blood pool. For the classification task in this study, both the segmented and non-segmented datasets were utilized. This dual approach allowed for a comprehensive analysis, leveraging the benefits of segmentation while also exploring the raw imaging data.

Figure 1 depicts the LGE modality with original slices and corresponding segmentation masks, Figure 2 shows Native T1 Mapping modality with Base, Mid, and Apex slices, and Figure 3 presents Post-Contrast T1 Mapping modality with Base, Mid, and Apex slices.

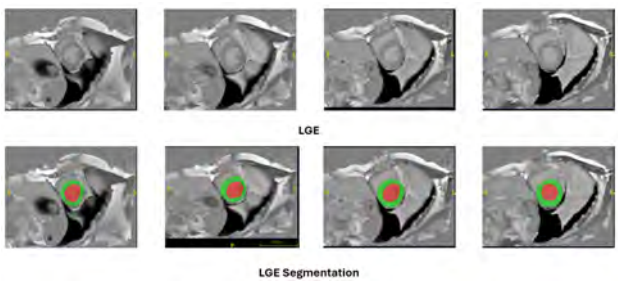



Figure 1: LGE modality with original slices and corresponding segmentation masks. The top row shows the original LGE slices, and the bottom row shows the LGE slices with segmentation masks.

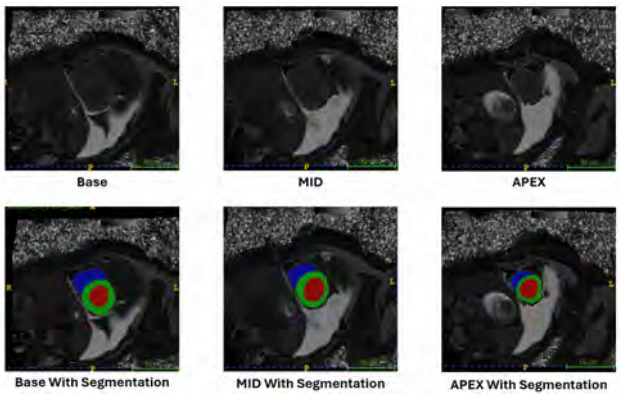



Figure 2: Native T1 Mapping modality with Base, Mid, and Apex slices. The top row shows the original slices, and the bottom row shows the slices with segmentation masks.

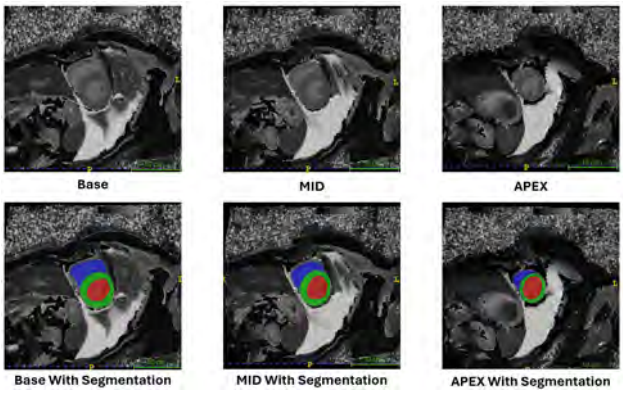



Figure 3: Post-Contrast T1 Mapping modality with Base, Mid, and Apex slices. The top row shows the original slices, and the bottom row shows the slices with segmentation masks.

### *3.2. Fusion Process*

The fusion process integrates volumetric data from three cardiac magnetic resonance (CMR) imaging modalities: Late Gadolinium Enhancement (LGE), native T1 mapping, and post-contrast T1 mapping. Each modality provides unique information about myocardial tissue characteristics. LGE: Highlights myocardial scar or fibrosis post-contrast, Native T1 Mapping: Quantitative T1 relaxation times pre-contrast, and Post-contrast T1 Mapping: T1 relaxation times post-contrast. Combining native T1 mapping and post-contrast T1 mapping proved particularly effective. But Including LGE did not yield additional benefits. The fusion algorithm spatially aligns and integrates features from native and post-contrast T1 mapping, producing a fused volume for classification. Figure 4 illustrates the fusion process, showing the integration of the three modalities into a comprehensive fused volume.

### *3.3. Dataset Split*

After the fusion process was performed, we started splitting the dataset to ensure a robust evaluation of the

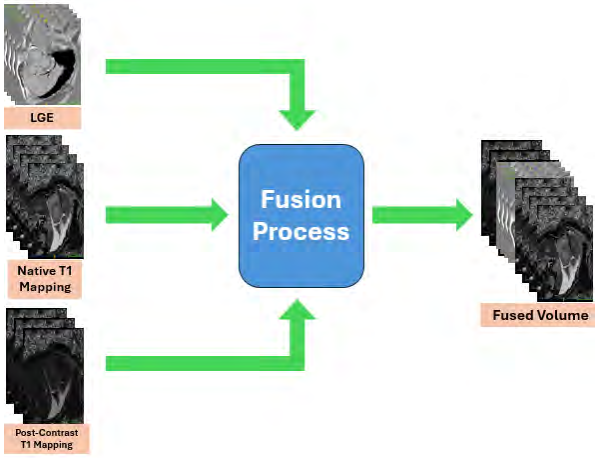



Figure 4: Fusion Process of CMR Modalities.

classification model, the dataset was divided using 5-fold cross-validation. Each fold was split into, Training Set: 70% of the data, Validation Set: 15% of the data, and Test Set: 15% of the data. The split was stratified to ensure that the distribution of each class was maintained equally across the folds. This stratification process also ensured that images from all three modalities (LGE, native T1 mapping, and post-contrast T1 mapping) for each patient were kept together, preventing data leakage between the training, validation, and test sets. A diagram illustrating the data split is shown in Figure 5.

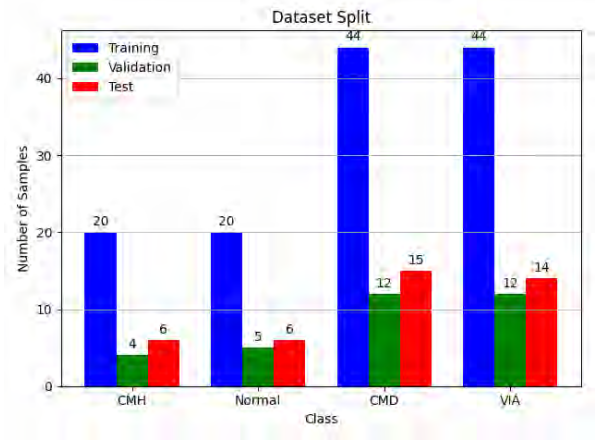



Figure 5: Showing how the dataset was split.

### 3.4. Classification methodology

Before explaining the final classification methodology, it is important to note that this method was arrived at after extensive experimentation with various classification techniques. The methodology that proved most effective is a Binary-Classification-Based approach for multi-pathology classification. This approach simplifies the complex task of multi-class classification by breaking it down into a series of binary classification problems, effectively overcoming class imbalance and the complexity associated with multiple pathologies. The process begins with four distinct cardiac pathologies: Cardiomyopathy (CMD), Viral Myocarditis (VIA), Hypertrophic Cardiomyopathy (CMH), and Normal. To streamline the classification, CMH and Normal cases were concatenated into one group, and CMD and VIA were concatenated into another group. This setup allowed us to feed these combined classes into the initial classification model.

#### 3.4.1. Initial Model (Model 1st):

The combined classes (CMH_Normal and CMD_VIA) are fed into the first model. This model performs the initial classification, predicting whether a sample belongs to the CMH_Normal group or the CMD_VIA group.

#### 3.4.2. Secondary Models:

Model 2nd: If the prediction from the first model is CMH_Normal, the sample is then fed into the second model, which further differentiates between CMH and Normal cases.
Model 3rd: If the prediction from the first model is CMD_VIA, the sample is then fed into the third model, which differentiates between CMD and VIA cases.

#### 3.4.3. Final Prediction:

The final classification result is determined by combining the outputs from the second and third models. This step-by-step binary classification approach ensures more accurate and reliable predictions for each pathology.
By using this methodology, we effectively address the issues of class imbalance and the complexity inherent in classifying multiple pathologies. The diagram 7 illustrates the step-by-step classification process, starting with the initial model (Model 1st) that distinguishes between combined classes (CMH_Normal and CMD_VIA). Depending on the initial prediction, the sample is then processed through either Model 2nd (for CMH vs. Normal) or Model 3rd (for CMD vs. VIA). The final classification result is derived from the combined outputs of the secondary models.

### 3.5. Main Experiments

As mentioned earlier, we utilized a fusion process and segmentation of the dataset from another project. The main experiments conducted in this study are as follows:

#### 3.5.1. Experiments without Segmentation

For the dataset without segmentation, the following experiments were conducted:

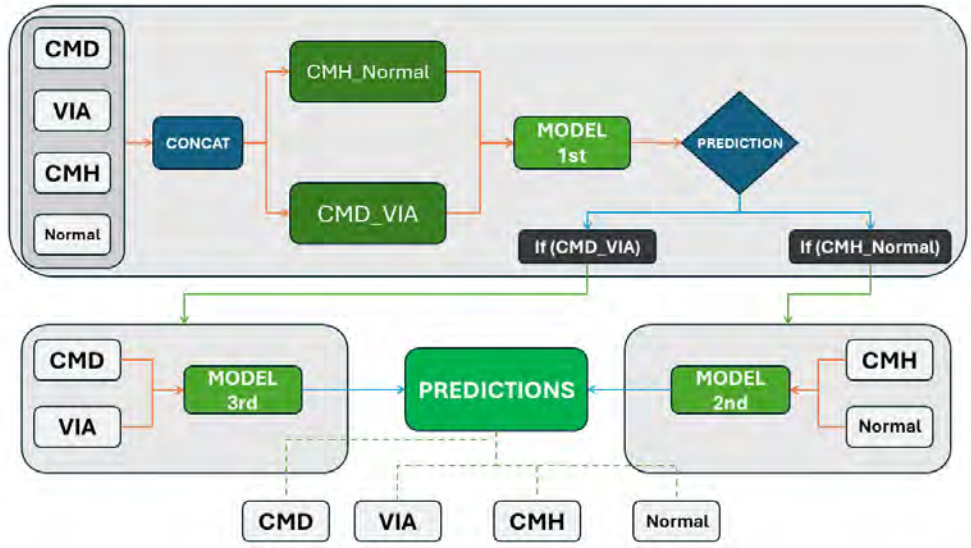



Figure 6: Binary-Classification-Based Method for Multi-Pathology Classification.

- **T1 Mapping (Native and Post Contrast)**: Fused data from native T1 mapping and post-contrast T1 mapping.
- **T1 Mapping Native**: Data from native T1 mapping alone.
- **T1 Mapping Post Contrast**: Data from post-contrast T1 mapping alone.
- **LGE**: Data from Late Gadolinium Enhancement (LGE) alone.
- **T1 Mapping (Native and Post Contrast) and LGE**: Fused data from native T1 mapping, post-contrast T1 mapping, and LGE.

These experiments were designed to test the classification performance using various combinations of T1 mapping and LGE modalities:

- **T1 Mapping (Native and Post Contrast)**: Evaluated the fused data from both native and post-contrast T1 mapping.
- **T1 Mapping Native**: Assessed the data from native T1 mapping independently.
- **T1 Mapping Post Contrast**: Assessed the data from post-contrast T1 mapping independently.
- **LGE**: Evaluated the data from LGE independently.
- **T1 Mapping (Native and Post Contrast) and LGE**: Assessed the fused data from native T1 mapping, post-contrast T1 mapping, and LGE.

### *3.5.2. Experiments with Segmentation*

The same set of experiments was repeated using the segmented dataset. The goal was to compare the classification performance with and without segmentation. The segmentation provided additional anatomical information, potentially enhancing the accuracy of the classification models.

### *3.5.3. Observations*

Upon conducting these experiments, it was observed that the classification performance using the segmented dataset significantly outperformed the dataset without segmentation. This result highlights the importance of segmentation in improving the accuracy and reliability of classification models.

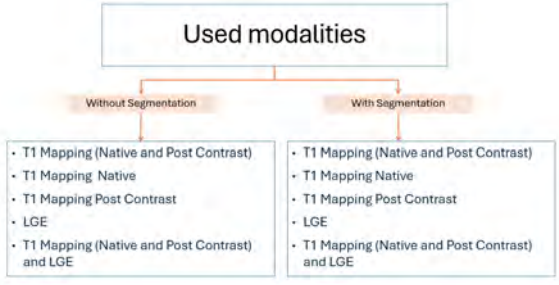



Figure 7: This figure illustrates the various experiments conducted in this study.

### *3.6. Classification Methodology Architecture*

Initially, we experimented with 3D model architectures to process the volumetric cardiac magnetic resonance (CMR) imaging data. However, the performance of the 3D models was poor. Consequently, we explored 2D approaches, which proved to be significantly more effective compared to the 3D approach.

*3.6.1. Without Segmentation*

The classification architecture for the dataset without segmentation is designed to process volumetric cardiac magnetic resonance (CMR) imaging data and predict the pathology. The process involves several key steps, which are illustrated in Figure 8.

*Input Volume.* The input to the model consists of volumetric CMR images. These volumes are divided into individual slices, each representing a cross-sectional view of the heart.

*Slices Extraction.* The input volume is split into multiple slices (s1, s2, s3, ..., last slice). Each slice is processed independently through a series of convolutional neural networks (CNNs).

*CNN Processing.* Each slice is fed into a separate CNN. After investigating various models such as ResNets and VGGs, EfficientNet B4 was selected for this task due to its superior performance.

- **EfficientNet B4**: Each slice is processed using EfficientNet B4, which extracts features through several convolutional layers and pooling layers. EfficientNet B4 was chosen for its balance between accuracy and computational efficiency, proving to be the best among the evaluated models.

*Feature Concatenation.* The features extracted from each slice by the CNNs are then concatenated to form a combined feature vector. This step aggregates information from all the slices, providing a comprehensive representation of the entire volumetric data.

*Fully Connected Layer.* The concatenated feature vector is passed through a fully connected layer. This layer further processes the combined features to generate a final feature representation suitable for classification.

*Pathology Prediction.* The output of the fully connected layer is used to predict the pathology. The model assigns a class label to the input volume, indicating the specific cardiac condition.

*3.6.2. With Segmentation*

In this approach, we utilized segmented data to enhance the classification performance and to investigate whether incorporating segmentation masks would improve the model's performance. The segmentation provides additional anatomical information, which is incorporated into the model along with the original CMR images. The process involves several key steps, as illustrated in Figure 9.

*Segmentation Input.* In addition to the original CMR slices, segmentation masks are included. The number of segmentation masks varies by modality:

- **LGE Modality**: Two masks are provided, highlighting the left ventricular blood pool and myocardium.
- **Native and Post-contrast T1 Mapping Modalities**: Three masks are provided, highlighting the left ventricular blood pool, myocardium, and right ventricular blood pool.

*Original and Segmentation Slices.* For each slice (s1, s2, s3, ..., last slice), both the original CMR image and the corresponding segmentation masks are combined to form multi-channel inputs:

- **LGE Modality**: The original slice (1 channel) is concatenated with the two segmentation masks (2 channels), resulting in a 3-channel input.
- **Native and Post-contrast T1 Mapping Modalities**: The original slice (1 channel) is concatenated with the three segmentation masks (3 channels), resulting in a 4-channel input.

*CNN Processing.* Each combined slice (original image + segmentation masks) is fed into a separate convolutional neural network (CNN). EfficientNet B4 was selected for this task due to its superior performance.

*Feature Aggregation and Prediction.* The features extracted from the CNNs encoders are then concatenated for all slices and segmentation masks. These combined features are then fed to the fully connected layer, yielding predictions in the same manner as the approach without segmentation.

*3.7. Uncertainty Quantification from All Experiments*

To enhance the reliability and robustness of our classification model, we incorporated an uncertainty quantification process. This process helps to evaluate the confidence of the model's predictions and improve the overall classification accuracy. The methodology is illustrated in Figure **??** and involves several key steps.

*Experiments.* The fused test volume is used as the input for multiple models trained on different main experiments. Specifically, the following main experiments are considered:

- **T1 Mapping (Native and Post Contrast)**
- **Native T1 Mapping**
- **T1 Mapping Post Contrast**
- **LGE**

Each of these experiments contributes to the overall prediction by generating individual model predictions. However, the last experiment combining Native and Post-contrast T1 mapping was skipped due to poor results obtained from it.

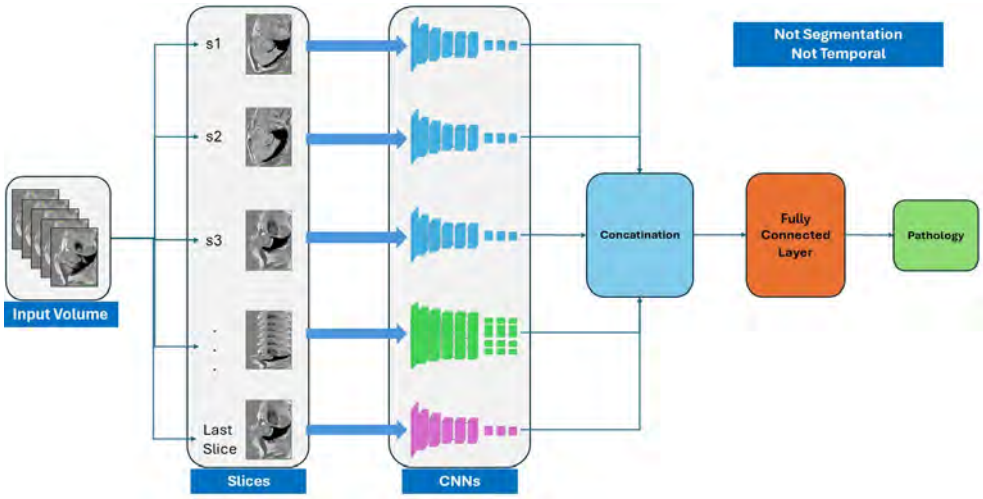



Figure 8: This figure illustrates the process from input volume through slices extraction, CNN processing with EfficientNet B4, feature concatenation, and final pathology prediction

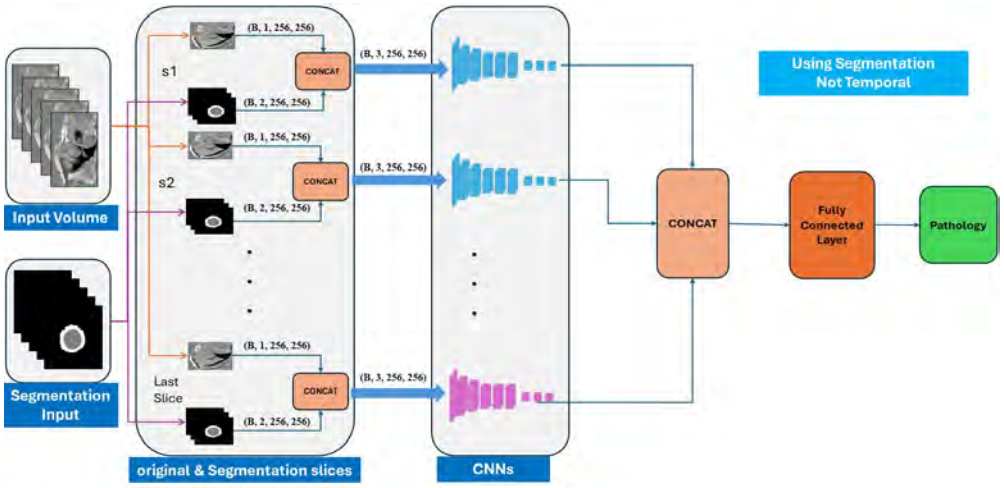



Figure 9: This figure illustrates the process from input volume and segmentation masks through slices extraction, CNN processing with EfficientNet B4, feature aggregation, and final pathology prediction

*Predictions.* For each test sample, the models trained on the different modalities provide separate predictions. These predictions represent the initial classification results based on the specific features extracted from each modality.

*Uncertainty Process.* The individual predictions from each model are then fed into an uncertainty quantification process. This process evaluates the confidence of each prediction, identifying potential uncertainties in the model outputs. The uncertainty process involves:

- Assessing the agreement among the different model predictions.
- Identifying predictions with high variability, which indicates uncertainty.

The uncertainty is quantified by majority voting from the predictions coming from the models.

*Voting.* To determine the final classification, a voting mechanism is applied. This mechanism aggregates the predictions from all models, taking into account the uncertainty scores. The final prediction is based on a consensus approach, where the most confident and consistent prediction is selected.

*Final Result.* The final result of the uncertainty quantification process is the predicted pathology. This result leverages the combined strength of multiple models and the uncertainty assessment to provide a more reliable and accurate classification.

### *3.8. Experimental Setup*

The model backbone used in our experiments is EfficientNetB4, which is well-suited for tasks involving

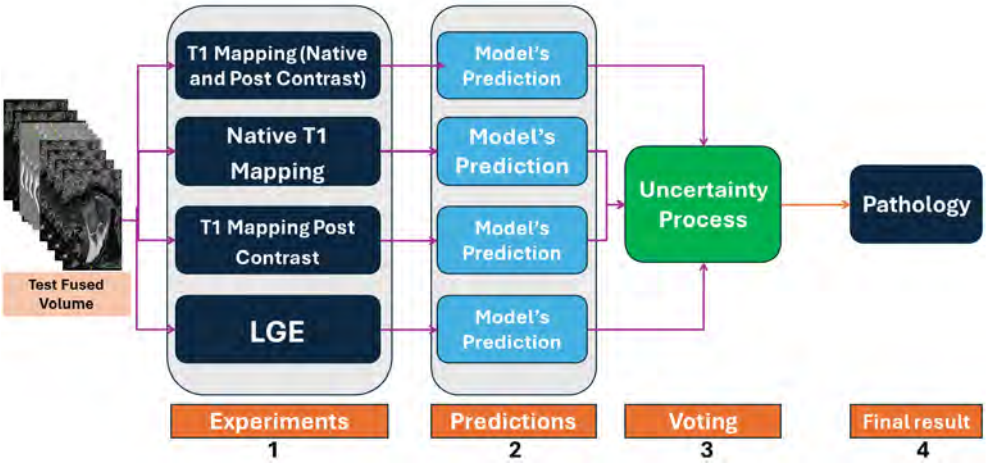



Figure 10: This figure illustrates the process from input volume and segmentation masks through slices extraction, CNN processing with EfficientNet B4, feature aggregation, and final pathology prediction

| Classes | Modality | Acc (Mean) | Acc (Std) |
|---|---|---|---|
| CMH vs. VIA | Native & Post Contrast T1 | 68.08 | 5.08 |
| | Native T1 | 73.08 | 5.17 |
| | Post Contrast T1 | 70.17 | 5.02 |
| | LGE | 78.00 | 5.33 |
| CMD_VIA vs. CMH_Normal | Native & Post Contrast T1 | 79.67 | 12.19 |
| | Native T1 | 78.70 | 7.21 |
| | Post Contrast T1 | 78.17 | 7.41 |
| | LGE | 74.78 | 5.78 |
| CMH vs. Normal | Native & Post Contrast T1 | 77.05 | 16.00 |
| | Native T1 | 75.51 | 5.03 |
| | Post Contrast T1 | 77.05 | 13.66 |
| | LGE | 69.23 | 17.47 |

Table 1: Mean and Standard Deviation of the Test Accuracy across the 5 folds for Each Experiment for the 3 binary classification problems

image data due to its balance of accuracy and computational efficiency. We train the model with a batch size of 4 over 50 epochs. The input to the model consists of 4 channels, with each image resized to a width and height of 220 pixels.

For optimization, we employ the Adam optimizer, which is known for its efficiency and ease of use. The learning rate is set to 0.0001, and a dropout rate of 0.2 is applied to prevent overfitting. The loss function used is Cross Entropy, which is standard for classification tasks.

We utilize a Tesla V100S-PCIE-32GB GPU with 32768MiB of memory for training, ensuring that our model can leverage high computational power for faster training times. The primary metric for evaluating the model's performance is accuracy, and we use the validation accuracy ('val_accuracy') as the criterion for model selection.

This setup ensures a robust training process, aimed at achieving high performance in the task of MRI cardiac pathology classification.

## 4. Results and Discussion

### *4.1. Challenges and Solutions*

During the course of this project, several significant challenges were encountered. Initially, one of the primary issues was dealing with the missing slices in the LGE volumes. To address this, various methods were explored, including reducing the number of slices in all volumes to match the volume with the fewest slices, padding slices to match the volume with the most slices, duplicating slices, and applying interpolation techniques to fill in the missing slices. Each of these methods had its own advantages and limitations, and through extensive experimentation, interpolation proved to be the most effective solution.

Another major challenge was the significant class imbalance present in the dataset. The initial approach to mitigate this involved reducing the number of samples in the majority classes to match those in the minority classes. While this helped, it did not fully resolve the

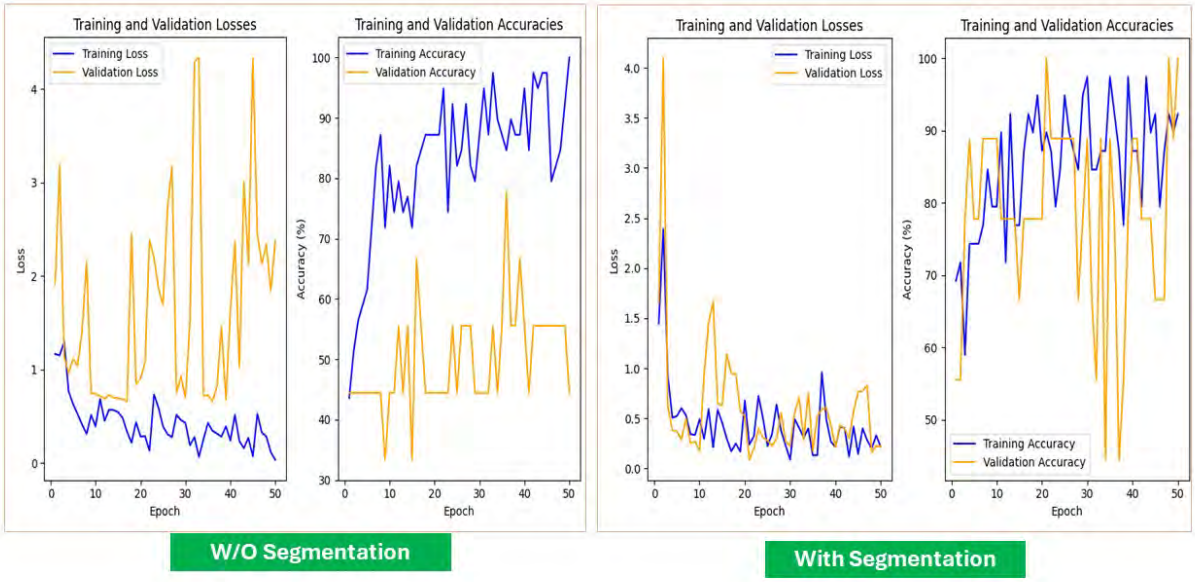



Figure 11: This figure contrasts the training and validation losses and accuracies for models trained with and without segmentation masks. This plot is generated from training and validation of the binary classification of (CMH vs. Normal) using the Native and Post Contrast T1 mapping modalities

| Classes | Acc (Mean) | Acc (Std) |
|---|---|---|
| CMD vs. VIA | 68.77 | 3.25 |
| CMD_VIA vs. CMH_Normal | 40.60 | 5.76 |
| CMH vs. Normal | 72.18 | 9.22 |

Table 2: Mean and Standard Deviation of the Test Accuracy across the 5 folds using the majority voting for the 4 experiments for the 3 binary classification problems

issue. Subsequently, a binary-classification-based approach was adopted to handle the multi-class classification problem. This method involved breaking down the multi-class problem into a series of binary classification tasks, which significantly improved the model's performance. This approach not only simplified the classification task but also enhanced the model's accuracy and robustness. These challenges required considerable time and effort to overcome.

### *4.2. Analysis of training with anatomical segmentation masks*

We initiated the experiments by training without segmentation masks (see Figure 8). However, we realized that there was a gap between the training and validation plots. Therefore, we decided to incorporate segmentation masks for the cardiac anatomy which may guide the model in the learning process (see Figure 9. We illustrate the training and validation plots in Figure 11.

#### *4.2.1. Without Segmentation*

The training loss decreases steadily, indicating that the model learns from the training data. However, the validation loss exhibits significant fluctuations and overall higher values compared to the training loss. This suggests that the model might be overfitting to the training data, failing to generalize well to unseen validation data.

The training accuracy steadily increases, reaching high values, but the validation accuracy fluctuates greatly. The high variance in validation accuracy, coupled with the divergence between training and validation accuracies, further supports the presence of overfitting.

#### *4.2.2. With Segmentation*

Both training and validation losses are lower and decrease more smoothly compared to the model trained without segmentation. The closer alignment between the training and validation loss curves suggests better generalization and less overfitting.

The training accuracy still increases steadily, but the validation accuracy is higher and fluctuates less compared to the model without segmentation. The reduced

gap and variability between training and validation accuracies indicate improved stability and generalization performance.

#### *4.2.3. Overall Analysis*

Training with segmentation masks leads to more stable and lower training and validation losses, as well as more consistent and higher validation accuracy. This suggests that incorporating segmentation masks helps the model to focus on the relevant anatomical features, improving its ability to generalize to new data and reducing overfitting. Consequently, using segmentation masks appears to be beneficial for enhancing the model's performance in MRI cardiac pathology classification tasks.

### *4.3. Binary classification problems*

Here, we start with discussing the 3 binary classification problems (see Table 1) independently as explained in Figure 9, and then we discuss the results of using the majority voting technique (see Table 2) as illustrated in Figure 10.

#### *4.3.1. CMH vs. VIA classification*

The results from the deep learning model for classifying cardiac pathologies (CMH and VIA) using different MRI modalities reveal several important trends. Firstly, it is evident that the LGE (Late Gadolinium Enhancement) modality achieves the highest mean accuracy at 78.00% with a standard deviation of 5.33%. This suggests that LGE images provide the most discriminative features for distinguishing between the two pathologies, which aligns with its known utility in highlighting fibrosis and scarring in cardiac tissues.

Following LGE, the Native T1 modality exhibits the second highest mean accuracy of 73.08% and a standard deviation of 5.17%. This result indicates that Native T1 imaging is also quite effective, albeit slightly less so than LGE, in identifying the pathologies. The higher accuracy compared to Post Contrast T1 and combined Native & Post Contrast T1 suggests that native tissue characteristics captured in T1 images play a significant role in pathology classification.

The Post Contrast T1 modality comes next with a mean accuracy of 70.17% and a standard deviation of 5.02%. This performance is slightly lower than that of Native T1, indicating that the contrast enhancement may not add significant value over the native images alone for this classification task. However, it still outperforms the combined Native & Post Contrast T1 modality. Interestingly, the combined Native & Post Contrast T1 modality has the lowest mean accuracy at 68.08% and a standard deviation of 5.08%. This outcome suggests that combining these modalities does not synergistically improve classification performance and may, in fact, introduce redundancy or noise that hinders the model's ability to accurately classify the pathologies.

In summary, LGE modality stands out as the most effective for this binary classification task, followed by Native T1 and Post Contrast T1, with the combination of Native and Post Contrast T1 being the least effective. These findings underscore the importance of selecting appropriate imaging modalities based on their individual contributions to the classification performance.

#### *4.3.2. CMH_Normal vs. CMD_VIA classification*

The results for the classification task distinguishing between Normal/CMH cases and VIA/CMD cases using various MRI modalities show notable trends. The combined Native & Post Contrast T1 modality achieves the highest mean accuracy at 79.67% with a relatively high standard deviation of 12.19%. This suggests that while this combination can be very effective, its performance may vary significantly across different folds, indicating potential variability in the features or model stability.

The Native T1 modality follows closely with a mean accuracy of 78.70% and a lower standard deviation of 7.21%. This indicates that Native T1 images alone are almost as effective as the combined modality but with more consistent performance across different test sets. Similarly, the Post Contrast T1 modality shows a comparable mean accuracy of 78.17% and a standard deviation of 7.41%, reinforcing that both native and post-contrast images individually provide strong features for this classification task.

Interestingly, the LGE modality, which was the most effective in the previous classification task, exhibits the lowest mean accuracy at 74.78% with a standard deviation of 5.78%. Although LGE is known for highlighting fibrosis and scarring, it appears to be less effective in distinguishing between the combined classes of Normal/CMH and VIA/CMD, possibly because these combined classes do not rely as heavily on the specific features captured by LGE imaging.

In summary, while the combined Native & Post Contrast T1 modality achieves the highest mean accuracy for this classification task, it also shows higher variability. Native T1 and Post Contrast T1 modalities individually provide nearly equivalent and more stable performance. The LGE modality, despite its effectiveness in the previous task, is less effective here, suggesting that the discriminative features for this classification problem are better captured by T1-based imaging rather than LGE. This emphasizes the need for modality-specific analysis based on the pathology and classification task at hand.

#### *4.3.3. CMH vs.Normal classification*

The classification results for distinguishing between normal cases and CMH pathology using different MRI modalities provide valuable insights. For this task, both

the combined Native & Post Contrast T1 modality and the Post Contrast T1 modality achieve the highest mean accuracy at 77.05%. However, they show considerable variability, with standard deviations of 16.00 and 13.66, respectively. This indicates that while these modalities can be highly effective, their performance is inconsistent across different folds, suggesting potential sensitivity to variations in the data or model.

The Native T1 modality shows a slightly lower mean accuracy of 75.51% but with a much lower standard deviation of 5.03%. This implies that Native T1 images provide a more consistent and reliable performance for this classification task, even if the peak accuracy is slightly less than that of the combined or post-contrast modalities.

On the other hand, the LGE modality has the lowest mean accuracy at 69.23% with the highest standard deviation of 17.47%. This suggests that LGE imaging, while useful in identifying fibrosis and scarring, may not be as effective in distinguishing normal cases from CMH pathology compared to T1-based imaging modalities. The high variability further indicates that LGE's performance is highly dependent on the specific dataset and folds used, making it less reliable for this particular classification task.

In summary, while the combined Native & Post Contrast T1 and Post Contrast T1 modalities achieve the highest mean accuracies, their performance variability suggests a need for caution. The Native T1 modality, with its more stable performance, appears to be a robust choice for distinguishing normal cases from CMH pathology. The relatively lower and more variable accuracy of the LGE modality highlights its limitations for this specific classification problem, emphasizing the importance of modality selection based on the nature of the classification task.

### *4.4. Majority voting of each binary problem*

#### *4.4.1. CMD vs. VIA classification problem*

The majority voting method results (see Tabel 2) in a mean accuracy of 68.77% with a standard deviation of 3.25%. This relatively low standard deviation suggests that the majority voting approach provides consistent performance across the different folds. While the mean accuracy is moderate, the stability of the results indicates that combining the outputs of different modalities through majority voting may help mitigate some of the variability seen in individual modality performances.

#### *4.4.2. CMD_VIA vs. CMH_Normal classification*

In the CMD_VIA vs. CMH_Normal classification task, the majority voting approach yields a mean accuracy of 40.60%, with a standard deviation of 5.76%. This significantly lower accuracy compared to individual modalities suggests that majority voting might not be effective for this specific classification problem. The complexity and possible overlap in the feature space of these combined classes could be a reason for the poor performance. Additionally, the higher standard deviation reflects considerable variability, indicating that majority voting struggles to provide stable results for this task.

#### *4.4.3. CMH vs. Normal classification*

For the CMH vs. Normal classification problem, majority voting achieves a mean accuracy of 72.18% with a standard deviation of 9.22%. This result shows that majority voting improves the performance compared to some individual modalities, highlighting its potential in leveraging complementary information from different sources. However, the relatively high standard deviation suggests variability in performance across folds, indicating that while majority voting can enhance accuracy, it may also introduce some inconsistency.

Overall, the majority voting approach shows mixed results. It provides consistent performance for CMD vs. VIA classification, struggles with CMD_VIA vs. CMH_Normal classification, and offers improved but variable results for CMH vs. Normal classification. These findings underscore the importance of understanding the specific characteristics and challenges of each classification task when selecting and combining modalities. The variability in performance, particularly in complex classification problems, suggests that further refinement or alternative ensemble strategies might be necessary to achieve more robust results.

### *4.5. Individual modality vs. Majority voting technique*

#### *4.5.1. CMD vs. VIA classification problem*

The majority voting technique yields a mean accuracy of 68.77% with a standard deviation of 3.25%, which is lower than the best-performing individual modality (LGE with 78.00% ± 5.33%). While majority voting provides a consistent performance with lower variability, it does not outperform the highest accuracy achieved by individual modalities.

#### *4.5.2. CMD_VIA vs. CMH_Normal classification*

The majority voting technique significantly underperforms with a mean accuracy of 40.60% compared to individual modalities, where the highest mean accuracy is 79.67% (Native & Post Contrast T1). This suggests that majority voting may not be suitable for this particular classification task, possibly due to the complex nature of combining CMD and VIA with CMH and Normal classes.

#### *4.5.3. CMH vs. Normal classification*

For the CMH vs. Normal classification, majority voting achieves a mean accuracy of 72.18% with a standard deviation of 9.22%. This performance is comparable to the individual modalities but does not surpass the highest individual accuracy of 77.05% (both Native & Post

Contrast T1 and Post Contrast T1). However, the variability in the majority voting results is lower, indicating more consistent performance.

#### *4.5.4. Consistency vs. Performance*

Majority voting tends to provide more consistent results with lower standard deviation across folds, as seen in the CMD vs. VIA and CMH vs. Normal classification tasks. However, it does not necessarily improve mean accuracy and, in some cases (like CMD_VIA vs. CMH_Normal), significantly underperforms compared to the best individual modality.

#### *4.5.5. Individual Modality Performance*

Individual modalities, particularly LGE, and combinations of T1, often achieve higher mean accuracies, though with greater variability in some cases. The choice of the best modality can be task-specific, highlighting the importance of modality selection based on the specific classification problem. While majority voting offers consistency, it may not always enhance overall performance. For optimal results, carefully selecting and possibly combining the best-performing individual modalities might be more effective, especially in complex classification tasks.

## 5. Conclusions

In this thesis project, we developed a robust classification framework that effectively integrates multimodal CMR imaging data and advanced deep-learning techniques to classify various cardiac pathologies. By incorporating segmentation masks and utilizing EfficientNet architectures, the framework significantly improves classification accuracy and reliability. The experiments demonstrated that providing detailed anatomical context is key to improving the classification task. This highlights the potential for improved patient outcomes through early and accurate disease detection. This research contributes to the ongoing efforts in automated cardiac diagnostics, offering a promising avenue for enhancing diagnostic precision and efficiency in clinical settings.

## 6. Future Work

Due to the limited time available for this project, we have not yet completed the integration of the three binary models into a single, unified model capable of handling four classes, as depicted in Figure 7. Our goal is to finalize this comprehensive model in the next few days, prior to the thesis presentation.

## Acknowledgments

This work was funded by the French National Research Agency (ANR) under reference ANR-19-CE45-0001-01-ACCECIT. Computations were performed using HPC resources from DNUM CCUB (Centre de Calcul de l'Université de Bourgogne) and the Mésocentre de Franche-Comté. Hsham Ngim acknowledges the support of an Erasmus+ scholarship provided by the European Commission. Special thanks to my supervisor, Stéphanie Bricq, for her extraordinary guidance and support throughout this project.

## References

Arega, T.W., Bricq, S., Meriaudeau, F., 2021. Leveraging uncertainty estimates to improve segmentation performance in cardiac mr, in: MICCAI UNSURE Workshop 2021, Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, Strasbourg, France. Hal-03349833.

Haaf, P., Garg, P., Messroghli, D.R., Broadbent, D.A., Greenwood, J.P., Plein, S., 2016. Cardiac t1 mapping and extracellular volume (ecv) in clinical practice: a comprehensive review. Journal of Cardiovascular Magnetic Resonance 18, 89. URL: https://doi.org/10.1186/s12968-016-0308-4, doi:10.1186/s12968-016-0308-4.

He, H., Garcia, E.A., 2019. Addressing class imbalance in deep learning models. Journal of Artificial Intelligence Research 15, 431–454.

He, J., Zhang, Y., Li, S., 2021. Attention-based fusion for multimodal cardiac disease classification. Medical Image Analysis 68, 101–110.

Johnson, J., Khoshgoftaar, T., Wald, R., 2019. Survey on deep learning with class imbalance. Journal of Big Data 6, 1–54. Available at https://journalofbigdata.springeropen.com/articles/10.1186/s40537

Liu, M., Gao, H., Sun, X., 2022. Hybrid cnn-rnn model for cardiac mri analysis. Journal of Cardiovascular Magnetic Resonance 24, 1–12.

Maddox, W., Izmailov, P., 2019. Evaluating model uncertainty in medical imaging. Advances in Neural Information Processing Systems 32, 1359–1370.

Petersen, S., Lee, A., 2020. Fusion of t1 mapping and lge for improved cardiac diagnosis. IEEE Journal of Biomedical and Health Informatics 24, 1203–1212.

Swathy, M., Saruladha, K., 2022. A comparative study of classification and prediction of cardio-vascular diseases (cvd) using machine learning and deep learning techniques. ICT Express 8, 109–116. URL: https://www.sciencedirect.com/science/article/pii/S24059595210011 doi:10.1016/j.icte.2021.08.021.

Tan, M., Le, Q.V., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. Proceedings of the International Conference on Machine Learning 36, 6105–6114.

WHO, W.H.O., 2023. Cardiovascular diseases (cvds). World Health Organization URL: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-d

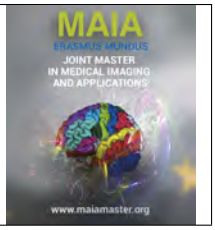
World Health Organization, 2017. Cardiovascular diseases (cvds). URL: https://www.who.int/newsroom/fact-sheets/detail/cardiovascular-di accessed: 2024-05-16.

Xinga, Y., Smith, J., Doe, A., 2024. Multimodal learning to improve cardiac late mechanical activation detection from cine mr images. arXiv preprint arXiv:2402.18507 URL: https://arxiv.org/abs/2402.18507.

Yuemeng Li, e.a., 2021. Acenet: Anatomical context-encoding network for neuroanatomy segmentation. Medical Image Analysis Available at https://doi.org/10.48550/arXiv.2002.05773.

Medical Imaging and Applications

Master Thesis, June 2024




# Enhancing the Prediction of Cognitive Decline by Integrating $^{18}$F-fluorodeoxyglucose Positron Emission Tomography ($^{18}$F-FDG PET) Radiomics and Clinical Variables Using Machine Learning

Andrew Dwi Permana*, Marco Bucci[a,c], Marina Bluma[a,c], Caroline Dartora[b,c]

[a]*Nordberg Translational Molecular Imaging Lab*
[b]*Westman Neuroimaging Group*
[c]*Center for Alzheimer Research, Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm, Sweden*

**Abstract**

As people age, cognitive functions such as memory, reasoning, language, concentrating, and processing speed can change. Cognitive impairment is possible for everyone and it can range from mild to severe. Mild cognitive impairment (MCI) is a condition where people can be characterized by a self-experienced loss of cognitive ability. MCI may not be severe enough to interfere with everyday living and independent functioning, as well as, there are no significant impairments on standardized cognitive tests. In contrast, dementia causes major difficulties in two or more domains of cognition, resulting in decreased independence. Due to the cognitive decline factor associated with MCI, there is well-documented evidence that some people with MCI will develop dementia, while others will not. For this reason, predicting the progression of MCI to dementia for early detection is becoming increasingly important, potentially slowing it down. Physicians have utilized brain imaging techniques, for instance, Positron Emission Tomography (PET) scans to assess neurodegenerative disorders. This technique provides useful information regarding the metabolism and physiology of the brain, including functional abnormalities. Recent studies have shown that $^{18}$F-fluorodeoxyglucose PET ($^{18}$F-FDG PET) can help improve the clinical diagnosis of people with MCI and dementia by revealing patterns of reduced glucose metabolism in the brain that are associated with neurodegenerative disorders. Moreover, neuroimaging currently widely uses the field of machine learning to improve the prediction of cognitive decline. In this study, we developed a predictive machine learning model that accurately predicts the progression of MCI to dementia by combining radiomics feature extraction and clinical variables from 277 $^{18}$F-FDG PET brain scans. For the experimental study, we investigated twelve different machine learning classification models with selected features using five different feature selection approaches on three different inputs: clinical variables only, radiomics features alone, and both data jointly. To evaluate our models' performance, we tested our prediction models using accuracy, F1-score, and Receiver Operating Characteristics - Area Under Curve (ROC - AUC). Our best model scored 0.83, 0.89, and 0.88 for accuracy, F1-Score, and AUC score, respectively. We obtained these outcomes by employing k-nearest neighbors (KNN) as a classifier and selected features using the ANOVA feature selection method.

*Keywords:* Dementia, Mild Cognitive Impairment (MCI), $^{18}$F-FDG PET, Machine learning, Radiomics, Predictive model

## 1. Introduction

Cognitive impairment is a situation when someone has difficulty learning and understanding new things, remembering, concentrating, and making decisions, which often occurs as a person ages (Centers for Disease Control and Prevention, 2018; Jessen et al., 2014). The condition of cognitive impairment can be mild to severe enough that it can interfere daily activities (Centers for Disease Control and Prevention, 2011). One until two percent of those 65 years of age or older have MCI, a common condition that affects a large portion of the senior population. From those people that are affected, around 10-15% will develop dementia each year

(Kumar et al., 2023). On the other hand, Alzheimer's Disease (AD) is a condition that occurs in individuals and is indicated by a progressive loss of cognitive and behavioural abilities (Corey-Bloom, 2003). Over 50 million individuals worldwide were diagnosed with dementia in 2020, according to the World Health Organization (WHO), and approximately 10 million new cases are reported each year and by 2050, there will be 139 million people, having nearly doubled every 20 years to reach 78 million in 2030 (Long et al., 2023). MCI is largely characterized by higher memory or thinking difficulties when compared to individuals of the same age; however, AD and other dementias are characterized by a wider range of symptoms that substantially interfere with daily activities (Centers for Disease Control and Prevention, 2011; Kumar et al., 2023; Langa and Levine, 2014). Thus, understanding the importance of early detection and intervention to potentially mitigate the progression to dementia is becoming essential (Langa and Levine, 2014).

Neuroimaging techniques have substantially improved the assessment of neurodegenerative disorders including dementia and MCI. One of the often-used techniques is $^{18}$FDG-PET. $^{18}$FDG is a glucose analog radiotracer that allows us to estimate glucose metabolism and biochemical activities in vivo in diseased and healthy tissues (Ashraf and Goyal, 2023). $^{18}$F-FDG PET scans are becoming useful for studying brain functionality and detecting abnormalities because the brain cells consume glucose for energy. When we want to assess the metabolic activity, in a healthy brain case, the regions with high neuronal activity will show high FDG uptake. In contrast, FDG uptake will reduce in certain brain regions, which indicates reduced metabolic activity and neuronal dysfunction (Minoshima et al., 2022). The region that is normally being assessed is the temporoparietal region, which includes parts of the temporal and parietal lobes for AD and in the case of frontotemporal dementia (FTD), the hypometabolism will show up in the frontal areas. These specific regions are significant in assessing MCI and dementia since they cover various cognitive functions. Furthermore, hypometabolism in the temporoparietal region is a crucial biomarker in assessing MCI and dementia, particularly AD. In people with MCI, this can also indicate a higher risk of progression to AD (Cerami et al., 2014).

In analyzing $^{18}$F-FDG PET images, it is often these images use color coding to measure different levels of $^{18}$F-FDG uptake, where the light colors reflect the high uptake areas (normal function activity), while the darker colors represent low uptake areas (hypometabolism). Moreover, a quantitative measure like Standardized Uptake Value (SUV) is commonly used to assess brain function, which normalizes $^{18}$F-FDG uptake by comparing it to the injected dose and the patient's body weight, the higher the SUV, it indicates the greater glucose metabolism (Ulaner, 2019).

Recent studies have been widely utilized $^{18}$F-FDG PET in both dementia research and clinical environments because it has the capability of accurately detecting changes in neuronal activity caused by neurodegeneration (Minoshima et al., 2022). Moreover, the European Academy of Neurology (EAN) and the European Association of Nuclear Medicine (EANM) have recommended the utilization of $^{18}$FDG-PET to enhance the clinical diagnosis of individuals with MCI, which may signify the initial phase of neurodegenerative disease, as well as those with dementia of unknown cause. Nevertheless, the existing literature on this subject is constrained in its formal evidently support (Chouliaras and O'Brien, 2023; Guedj et al., 2022).

The field of Artificial Intelligence (AI), in particular, Machine Learning methods, has gained significant developments in the medical field, specifically in the areas of image-based disease diagnosis, prognosis, and risk assessment (Chan et al., 2020; Cheplygina et al., 2019; Varoquaux and Cheplygina, 2022). These approaches have demonstrated the ability to analyze thousands of images in minutes and have shown performance comparable to that of trained physicians and radiologists (Cheplygina et al., 2019). The trend is increasingly toward using these techniques to process large amounts of data in a reliable manner. The $^{18}$F-FDG PET images extensively apply machine learning techniques to predict neurodegenerative diseases. This allows us to get important insights into cognitive progression and patient outcomes (Litjens et al., 2017; Rana and Bhushan, 2022). Machine learning algorithms have great potential for enhancing the precision of diagnosing and predicting the conversion of AD and MCI when used with $^{18}$F-FDG PET imaging due to their ability to do complex analysis (Rasi et al., 2024).

However, the traditional use of basic imaging characteristics in $^{18}$F-FDG PET imaging for disease classification has been associated with challenges in achieving satisfactory classification accuracy and clinical relevance. Fortunately, recent research that implements a new feature extraction method known as radiomics has demonstrated substantial prospects for overcoming these difficulties (Kumar et al., 2012). Radiomics is a rapidly evolving field in medical imaging that involves the extraction and analysis of a large number of quantitative features from medical images. Using the PyRadiomics platform, radiomics data can be extracted from different neuroimaging techniques, such as computed tomography (CT), PET, and magnetic resonance imaging (MRI) scans. The platform does this in four main steps: (i) loading and preprocessing the image and segmenting the maps; (ii) applying enabled filters; (iii) computing features using different feature classes; and (iv) showing the results (van Griethuysen et al., 2017).

In this study, we wanted to construct an integrated machine learning model that could predict cognitive decline, particularly the conversion of MCI to dementia,

by using the features that we extracted from radiomics and clinical variables from the $^{18}$FDG-PET brain images. By doing so, we expected to enhance the prediction of cognitive decline, which can save time, improve accuracy, and lead to better clinical outcomes.

## 2. State of the art

Cognitive decline is a major public health concern, as highlighted in the research. Subjective Cognitive Decline (SCD) refers to the personal perception of a decline or increased occurrence of confusion or memory loss. It is typically one of the first obvious signs of AD and other associated forms of dementia (Centers for Disease Control and Prevention, 2018). Predicting the progression of cognitive decline becomes an essential stage for improving early dementia detection and better strategic patient care. Dementia is a general term used to characterized by one or more neurodegenerative disorders resulting in a substantial decline in cognition that is significant enough to disrupt daily functioning (Dave et al., 2020). $^{18}$F-FDG PET imaging has been scientifically proven to be an effective tool for quantifying brain glucose metabolism that can lead to identifying the MCI and dementias through the amount of $^{18}$F-FDG uptake in the specific areas (Teng et al., 2020).

SUVR (Standardized Uptake Value Ratio) is considered the most common quantitative method used to quantify the glucose metabolism in specific brain regions that can help identify metabolic deficits (Vemuri et al., 2016). While SUVR and other techniques are helpful in assessing and analyzing $^{18}$F-FDG PET scans, there are some challenges that may arise. For instance, in traditional radiology practice, except for a few measurements like size and volume, the imaging data sets are typically assessed through visual or qualitative analysis. In addition to involving intra- and interobserver variability, this method may disregard a significant amount of hidden data within the medical images (Koçak et al., 2019). These conventional metrics, while widely used, it may not fully reflect all the available information, thus potentially limiting their utility in comprehensive disease characterization (Tixier et al., 2016).

Due to this, radiomics can contribute to tackling those challenges. Radiomics has the capability of collecting and organizing large amounts of data, which makes it highly suitable for studying complicated diseases with various aspects. As a result, it has mostly been researched in the field of oncology (Bevilacqua et al., 2023). Radiomics can extract information based on size, shape, borders, and heterogeneity (van Griethuysen et al., 2017). $^{18}$F-FDG PET scans can provide information for feature extraction, such as the intensity features will measure features from the distribution of voxel intensities, which indicates the concentration of $^{18}$F-FDG uptake in the brain. The geometry and size of the structural characteristics of ROI will contribute to its shape-based features. The texture feature assesses the spatial arrangement of voxel intensities, which provide information on heterogeneity and patterns of $^{18}$F-FDG uptake.

The main objective of radiomics is to extract as much as possible of useful hidden objective data that may be used for decision support (Koçak et al., 2019). Radiomics allows not only to extract of information using an original image as an input but also applying filters such as wavelet, Laplacian of Gaussian (LoG), LocalBinaryPattern3D and others. Moreover, after defining the image type, the next thing is to define the feature class, which defines the class where the features will be extracted. These feature classes include for example, First Order Statistics, Shape-based (2D and 3D), and Gray Level Co-occurrence Matrix (van Griethuysen et al., 2017).

In recent times, researchers have employed various methods to conduct innovative studies in this field, such as integrating medical image analysis with AI to enhance diagnosis, prognosis, and clinical outcomes (Castiglioni et al., 2021). Different studies (Feng et al., 2021; Li et al., 2019; Singh et al., 2023) have shown promising results in predicting dementia, particularly AD, using both MRI and $^{18}$F-FDG PET with radiomics features. The use of AI for analyzing the radiomics features is highly advantageous since it can handle the amount of data extracted better than traditional statistical methods (Koçak et al., 2019).

Moreover, in several studies, the progression of cognitive decline was predicted, for instance, in a recent study by Peng et al. (2023), the authors built a machine learning model that predicts progression from MCI to AD using white-matter and radiomics features. The authors analyzed $^{18}$F-FDG PET-based radiomics features from 341 MCI patients from the Alzheimer's Disease Neuroimaging Initiative (ADNI), of which 102 of them progressed to AD. The authors extracted the features from the white matter and dimensionally reduced them to construct a psychoradiomics signature (PS), and they combined them with multimodal data to build an integrated model. To evaluate the model performance, the authors used the ROC curves in the test group, with a score of 0.865. In the study of Shu et al. (2021), that used ADNI database consists of 357 patients with MCI, of whom 154 progressed to AD during the 48-month follow-up period. The authors aimed to use machine learning to build and validate a radiomics-integrated model based on brain MRI to predict MCI patients' conversion to AD. The integrated model based on whole-brain radiomics could accurately identify and predict the high-risk population of MCI patients who may progress to AD. The ROC curve showed that the accuracy of the model in the training and test sets was 0.814 and 0.807, respectively, with a progression to AD within 12 months.

Another study analyzed $^{18}$F-FDG PET/CT brain images to predict AD in patients with Amyloid PET Positivity. Pyradiomics was used to extract the feature and a machine learning algorithm, specifically discriminant analysis, was used to obtain the best diagnostic performance in the prediction. A total of 11 radiomics features that were important were selected. As regards the performance in the prediction of the final clinical-instrumental diagnosis of AD, the highest score of AUC with accuracy was 0.79 (Alongi et al., 2022).

There has been much research for utilizing the radiomics feature extraction technique based on MRI brain images compared to $^{18}$F-FDG PET (Chaddad et al., 2018; Feng et al., 2018; Li et al., 2018). Moreover, the study mostly focused on the progression from MCI to AD. For these reasons, our study mainly focused on applying a machine learning predictive model that employed radiomics feature extraction to investigate the 277 $^{18}$F-FDG PET brain image dataset to improve the prediction of cognitive decline from MCI to dementia by integrating the clinical variables and radiomics features and evaluating them using twelve different classifiers and five different feature selection methods.

## 3. Material and methods

### *3.1. Dataset*

In this work, the sample comprised 277 individuals who visited the Clinic for Cognitive Disorders, Theme Aging, at Karolinska University Hospital in Stockholm, Sweden. Among them, 177 individuals received a diagnosis of MCI, while the remaining 100 were diagnosed with dementia during the initial assessment. $^{18}$F-FDG PET data was processed at the Nordberg Translational Molecular Imaging Lab. The clinical diagnosis was assessed in three different visits by dementia experts. The first visit corresponded to the baseline clinical diagnosis. A second assessment was performed, including $^{18}$F-FDG PET results, biomarkers, neuropsychological testing, and additional information following the initial diagnosis. The third assessment was performed at an average of 4.7 months following the baseline diagnosis, it is also represented as the final follow-up diagnosis (Perini et al., 2021). In this study, we were focused on the $^{18}$F-FDG PET analysis only, which means we did not consider the baseline diagnosis, instead, our initial diagnosis is the post-$^{18}$F-FDG PET diagnosis. Our main research goal is to figure out how cognitive decline from MCI to dementia will progress. To do this, we divided the diagnosis into two groups: MCI group and dementia group. Patients diagnosed with memory syndrome (MS) and MCI were put together in the "MCI" group, this grouping was according to the assessment of Perini et al. (2021), based on our understanding, they were pooled together because of the similarities in cognitive decline and potential progression to dementia. The remaining patients with other diseases were put into the "Dementia" group. Table 1 describes the grouping of the diagnosis.

Table 1: Grouping of the diagnosis

| **Diagnosis** | **Group** |
|---|---|
| MCI | **MCI** |
| Memory syndrome | |
| Alzheimer's disease | **Dementia** |
| Amyotrophic Lateral Sclerosis | |
| Corticobasal degeneration | |
| Dem | |
| Dementia with Lewy bodies | |
| FTD | |
| lvPPA | |
| Posterior corticol atrophy | |
| Parkinson disease dementia | |
| Pick's disease | |
| PNFA | |
| Parkinson syndrome | |
| Progressive supranuclear palsy | |
| Psychiatric syndrome | |
| Spinocerebellar ataxia | |
| Vascular dementia | |

*MCI*: Mild Cognitive Impairment, *ALS*: Amyotrophic Lateral Sclerosis, *Dem*: Dementia not otherwise specified, *FTD*: Behavioral variant of Frontotemporal dementia, *lvPPA*: Logopenic variant of primary progressive aphasia, *PNFA*: Nonfluent variant of progressive aphasia

#### *3.1.1. $^{18}$F-FDG PET acquisition and analysis*

The $^{18}$F-FDG PET examinations were conducted at the Department of Medical Radiation Physics and Nuclear Medicine Imaging, Karolinska University Hospital, Stockholm, Sweden, utilizing a Biograph mCT PET/CT scanner (Siemens/CTI, Knoxville, TN). $^{18}$F-FDG PET was performed at an average of 4.7 ± 6.0 (mean ± SD) months following the baseline diagnosis (4.2 ± 4.3 months and 5.5 ± 8.4 months for patients diagnosed with MCI and dementia at baseline, respectively). During the process, all the patients were examined with open eyes in a 10-min or 15-min list-mode scan starting 30 to 45 min after intravenous injection of 2 – 3 MBq/kg (Perini et al., 2021).

#### *3.1.2. Patient characteristics at post-$^{18}$F-FDG PET and follow-up diagnosis*

The post-$^{18}$F-FDG PET diagnosis (grouped diagnosis) showed that the number of cases of dementia was 156 and 121 cases belonged to MCI. Both cases of dementia and MCI had similar mean follow-up times, which were (3.5 ± 1.8 and 3.8 ± 1.8 years, respectively) (Perini et al., 2021). At follow-up, 44 out of 121 MCI subjects (36%) had developed dementia, which made our dataset 77 cases in MCI and the rest of the 200 cases were dementia.

#### *3.1.3. Clinical variables*

Besides the $^{18}$F-FDG PET brain images dataset, as the purpose of this project is to integrate the clinical variables data with the radiomics feature, we analyzed and selected the clinical variables that can be useful when we combine them. To handle missing values in the biomarkers dataset, we used the scikit-learn machine learning library, which is the *KNNImputer* class that supports nearest neighbor imputation k-Nearest Neighbors. The algorithm finds the k nearest data points (neighbors) based on similarity in feature space, then the imputed value is computed by averaging (or weighted average) the values of the nearest neighbors and finally, the imputed value is replaced by the missing value. The second technique is Multiple Imputation by Chained Equations (MICE), which is a process based on the relationships between variables and estimates them to impute the missing value iteratively. Table 2 summarizes the clinical variables used and Table 3 describes the missing values.

Table 2: Demographic of clinical variables

| | Dementia (n=156) | MCI (n=121) |
|---|---|---|
| **Age**, mean (sd), years | 67.01 (8.71) | 64.25 (10.39) |
| **Gender**, *N.* (%) | | |
| Female | 75 (48) | 69 (57) |
| Male | 81 (52) | 52 (43) |
| **ttau**, *N.* (%) | | |
| Positive | 68 (49) | 37 (37) |
| Negative | 71 (51) | 62 (63) |
| **Ptau**, *N.* (%) | | |
| Positive | 35 (25) | 21 (21) |
| Negative | 104 (75) | 78 (79) |
| **mta**, *N.* (%) | 136 (56) | 107 (44) |
| **mmse**, mean (sd) | 23.06 (5.08) | 26.96 (2.64) |
| **gca**, *N.* (%) | 117 (56) | 93 (44) |
| **amyloid**, *N.* (%) | | |
| Positive | 43 (31) | 20 (19) |
| Negative | 94 (69) | 83 (81) |

*ttau*: total tau, *pTau*: Phosphorylated Tau, *mta*: medial temporal atrophy, *mmse*: mini mental state examination, *gca*: global cortical atrophy, *amyloid*: amyloid $\beta$ and amyloid PET

Table 3: The number of missing values in the clinical variables

| Clinical variable | Amount |
|---|---|
| **ttau** | 39 |
| **Ptau** | 39 |
| **mta** | 34 |
| **mmse** | 20 |
| **gca** | 67 |
| **amyloid** | 37 |

### *3.2. Working pipeline*

In this work, we incorporated radiomics feature extraction to extract features from the 277 $^{18}$F-FDG PET brain images. Starting with the input from the preprocessed $^{18}$F-FDG PET images, applying feature extraction, selecting the best features, and feeding them to the classifiers to predict the progression of cognitive decline. For more detail of the pipeline in this work, it is illustrated in Figure 1.

#### *3.2.1. Preprocessing $^{18}$F-FDG PET image*

The whole $^{18}$F-FDG PET image has been normalized to MNI space and smoothed. The preprocessing of $^{18}$F-FDG PET scan was done using global mean normalization. It started with the reconstruction of the raw image and was followed by motion correction to align it. The image was then spatially normalized to standard MNI anatomical space and was smoothed using a Gaussian filter in order to improve the signal quality. Then, the mean uptake value within the reference region in the brain (cerebellum) was calculated. After that, to obtain the SUVR, the PET scan uptake value of each voxel is divided by the mean reference region value. These preprocessing image processes were performed using SPM and MATLAB software.

#### *3.2.2. Grey mask normalization*

Elastix and Transformix approaches were utilized for the image spatial and intensity normalization of the preprocessed images to the MNI template to get the grey matter (GM) mask. Elastix is widely used to perform medical image registration (Klein et al., 2010). Elastix offers a range of registration algorithms that are suited to different image types and registration tasks. In this study, the algorithm we chose is affine registration. Since Elastix requires a parameter file that specifies the registration settings, we defined the parameter provided which is the *Parameter9 affine* (Artaechevarria et al., 2009), (details of the setting are provided under the appendix section). The radiomics extracted many numerical features, where the input, the radiomics needed the preprocessed $^{18}$F-FDG PET scan and its mask, in this case, the grey matter mask. To obtain the GM mask, we performed Elastix and Transformix to get the normalized image and registered label. We executed the normalization process using Elastix, and once it was completed, we obtained the transformation parameter that we used to perform the Transformix. The transformix process took the template label and the transformation parameter that obtained from the Elastix process. After the normalized and registered labelled images were created, the GM mask was then generated. Its process involved by creating a binary GM mask of the image from the registered label then applied to the original images. Figure 2 shows how the GM mask was generated.

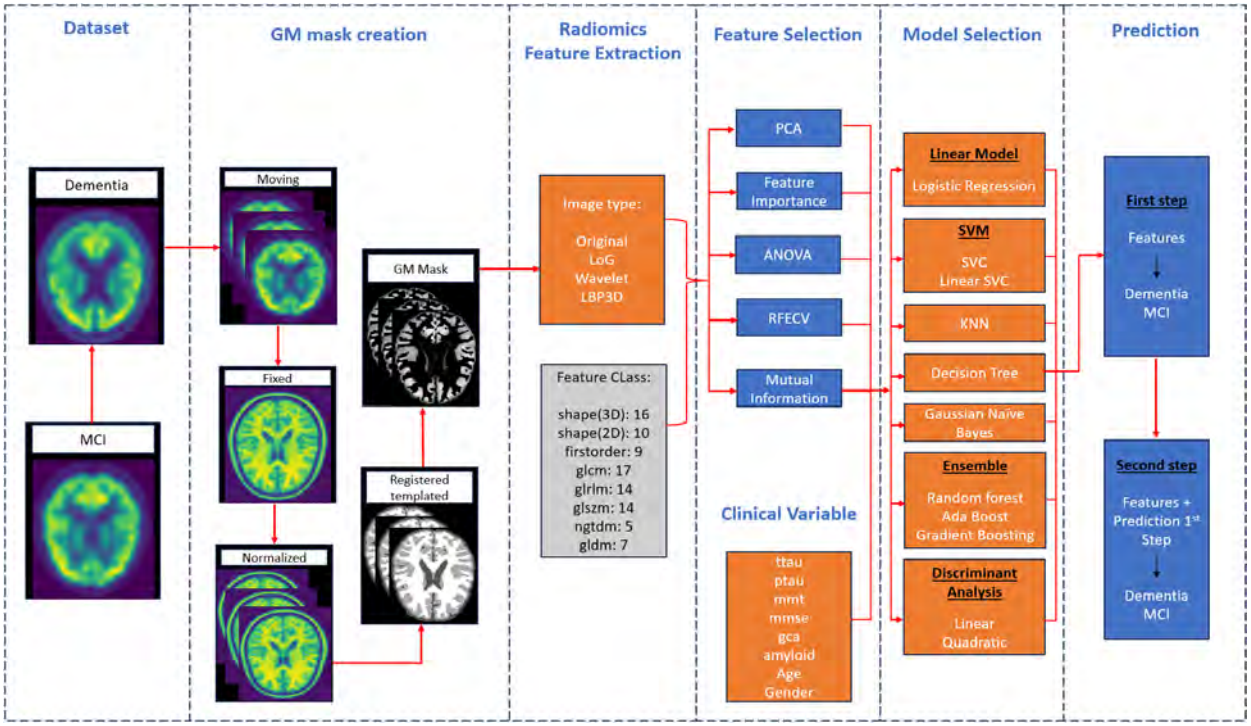



Figure 1: The pipeline for this work, starting with the GM mask creation, radiomics feature extraction, applying feature selection, and continuing to predict the cognitive decline progression with several different classifiers

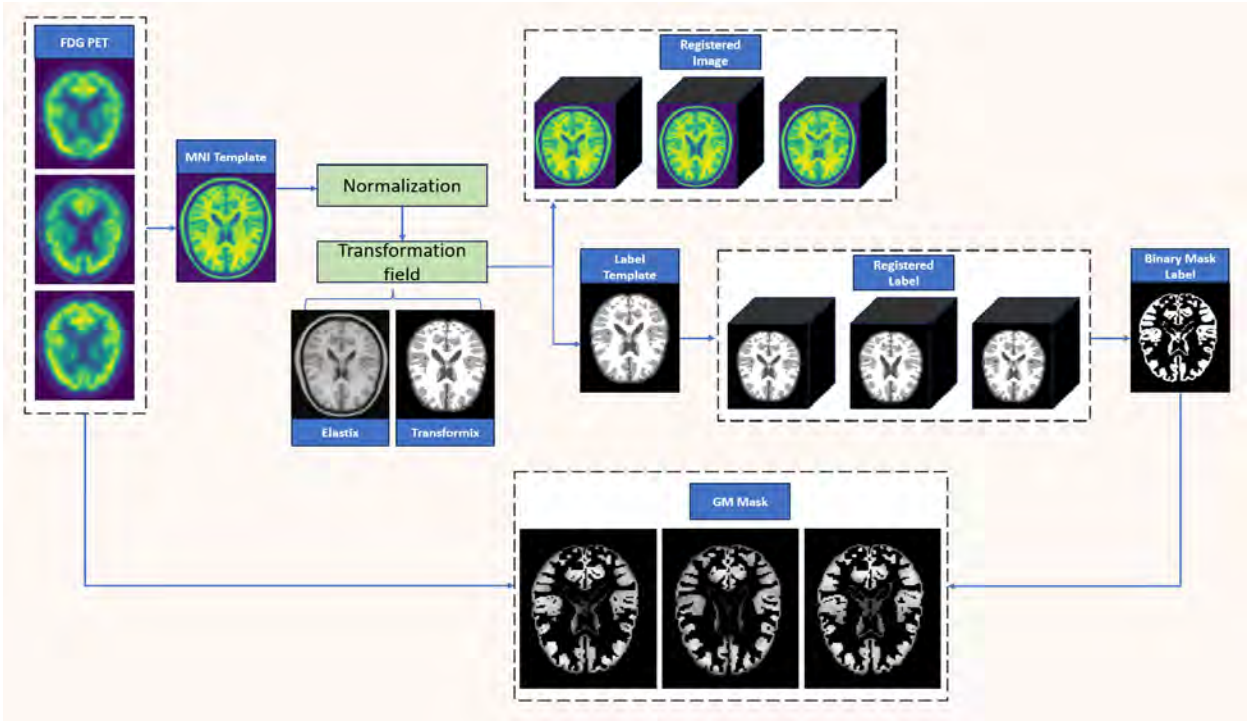



Figure 2: The pipeline for creating the GM mask, after the normalized image and registered label were created, the next step was to create a binary mask and finally the GM mask

### *3.2.3. Radiomics feature extraction*

Radiomics is a technique that involves the quantitative analysis of medical images that are commonly utilized in conventional medical practice. The process involves extracting a diverse array of manually designed features from medical images. The aforementioned variables are thereafter analyzed to ascertain their correlation with the prognosis and characteristics of patients

(Mannil et al., 2018; Rasi et al., 2024). Radiomic data are mineable, meaning that in such large datasets, it may be utilized to discover and identify previously unknown markers and patterns of the disease evolution, progression, and treatment strategies, thus, by utilizing radiomics, we expected to explore information from medical images and complex patterns that are challenging to identify and quantify using human eye (Mayerhoefer et al., 2020).

The Pyradiomics framework has different image types and feature classes that we can define for feature extraction (van Griethuysen et al., 2017). In this study, we calculated 92 features for each input in total. The Pyradiomics framework allowed us to define the input not only from the original image but also could apply different filters. We defined four different image types as input, the first one is the original image and the rest of three are images that were applied with filters, namely, LocalBinaryPattern3D (LBP3D), which computes the local binary pattern in 3D using spherical harmonics, Laplacian of Gaussian filter (LoG), known as edge enhancement, that highlights areas with changes in gray levels, with the degree of texture enhancement determined by the value of sigma, and Wavelet filtering, it yields 8 decompositions per level (all possible combinations of applying either a High or a Low pass filter in each of the three dimensions (van Griethuysen et al., 2017).

The extracted features were classified into various classes, which in this work were Shape-based(3D) (16 features), Shape-based(2D) (10 features), First Order (9 features), Gray Level Co-occurrence Matrix (GLCM) (17 features), Gray Level Run Length Matrix (GLRLM) (14 features), Gray Level Size Zone Matrix (GLSZM) (14 features), Neighbouring Gray Tone Difference Matrix (NGTDM) (5 features), and Gray Level Dependence Matrix (GLDM) (7 features) (van Griethuysen et al., 2017). Firstly, we defined each image type separately, and then we performed the radiomics technique to extract the feature, after 4 different image types had been extracted separately, the next thing, we defined in our algorithm for all the image types and extracted them together. This approached allow us to make an analysis to individual image types and the combination of all image types for a better comparison. All these settings, we defined them in the parameter file. For details of the parameter file, it can be seen in the appendix section. Figure 3 illustrates the overview of PyRadiomics framework process.

### *3.2.4. Feature selection*

While radiomics extracts a huge amount of features, it frequently tends to redundancy and irrelevant information which may result in overfitting during the classification phase. Before we performed feature selection, we did preprocessing that involved discarding constant, quasi-constant, and duplicate features manually. Per-

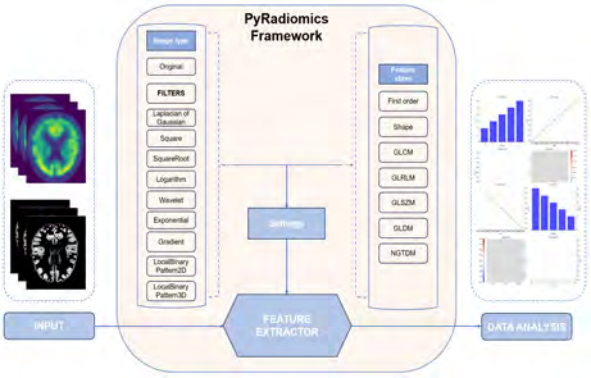



Figure 3: Overview of PyRadiomics process, starting from the input images which are the original image and its mask, then applying the defined settings, and performed the feature extraction

forming feature selection to select the most relevant features that can optimize the classification task is essentially important and it makes the model perform well. In this work, we utilized five different feature selection techniques, specifically:

- *Principal Component Analysis (PCA)*
  PCA has been used in machine learning for feature extraction based on dimensionality reduction of high-dimensional imaging. The algorithm selects only the most significant components to retain, while the insignificant ones are eliminated. PCA creates the new features that are a linear combination of the original attributes and vectors, in a dataset with $d$ dimensions, PCA reduces it to a $k$-dimensional space where $k$ is less than $d$. These new features, called principal components (PCs), and each PC captures the maximum amount of variance while excluding other sources of variance. (Sudharsan and Thailambal, 2023). In this work, we used the *n component*: 0.9.

- *Feature Importance (FI)*
  The concept of the feature importance is calculating the increase from the error that obtained in the prediction model after permuting the feature. A feature is considered ”important” when the model relied on the feature for the prediction because the values increase the model error. On the other hand, an ”unimportant” feature does not change the model error, because in this case, the model ignores the feature for the prediction (Molnar, 2019). Random forest method was utilized for the classifier in this approach as the estimator and the number of the estimator that we wanted to select was 10 features.

- *Analysis of Variance (ANOVA)*
  ANOVA is a set of statistical models and estimation processes to evaluate if the means of two or more data samples are from the same distribution.

The ANOVA is a type of F-test also known as F-statistic, it is an univariate statistical test where each feature is compared to the target variable and evaluated if the feature have a statistically significant relationship. ANOVA is usually used in classification tasks when the input data are numerical and the target feature is categorical (Pathan et al., 2022). ANOVA can help identify which radiomics features have significant differences between the categories, this can help with feature selection and improving model performance.

- *Recursive Feature Elimination with cross validation (RFECV)*
  RFECV is a wrapper feature selection technique that utilize machine learning algorithm in order to find and select the most relevant features. This technique utilize cross validation aiming to make sure its robustness and reliability when determining the most suitable amount of features that optimizes the performance of the model (Awad and Fraihat, 2023). RFECV uses a classification machine learning model, in this work, SVC classifier to evaluate and assign a score to each feature. In each iteration, it eliminates features that do not contribute to improve the accuracy of the classification.

- *Mutual Information (MI)*
  Mutual Information has been commonly utilized in machine learning for feature selection, it is based on a filter method to measure relevance and redundancy for selecting optimal features in predicting the target variable with respect to other variables (Beraha et al., 2019). The algorithm is trying to identify a subset of features that is showing a strong correlation (high mutual information) with the target variable, while also minimizing redundancy (low mutual information) among the selected features.

#### *3.2.5. Classification*

In this study, our main objective is to predict the progression of cognitive decline by classifying between dementia and MCI. To achieve this aim, we proposed a two-step hierarchical classification, where the first step was contributed to the second step as a time series classification, meaning that, the features and the prediction from the first step were used in the second step to predict the progression. In the first step, we predict the initial diagnosis (grouped diagnosis) between dementia and MCI, then in the second step, we used the follow-up diagnosis as a reference to predict the cognitive decline from MCI to dementia. We utilized 12 different classifier techniques, in detail, Support Vector Classifier (SVC), Linear SVC, Random Forest (RF), KNN, Decision Tree (DT), Linear Regression (LR), Ada Boost (AB), Gradient Boosting (GB), Linear Discriminant Analytics (LDA), LDA with shrinkage, Quadratic Discriminant Analytics (QDA), and Gaussian Naive Bayes (GNB). By choosing all these different classifiers, we wanted to explore and evaluate a range of machine learning modelling techniques to find the most effective, accurate, and robust solution for our classification and prediction tasks. This approach helped in making informed decisions based on empirical performance and understanding the strengths and limitations of various algorithms.

#### *3.2.6. Data augmentation*

During the process of our work, we encountered a challenge regarding the dataset, since we grouped the neurodegenerative diseases into dementia and MCI, in the second step, the dataset became imbalanced, where the distribution of dementia was 200 cases and the MCI only 77 cases. To handle this imbalanced dataset, we tried to do a data augmentation by oversampling the dataset to help improve the performance and generalization ability of our machine learning models. We employed two different data augmentation techniques, including the SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling). The concept of SMOTE is to balance the dataset by generating synthetic samples for the minority class, reducing the risk of overfitting that might occur. SMOTE has been improved by ADASYN, with a focus on those minority class samples that are harder to classify (Yakshit et al., 2022).

#### *3.2.7. Evaluation performance*

- *Nested cross validation*
  A cross validation (CV) is an essential step in developing a machine learning algorithm since it estimates its prediction error. Cross validation also can effectively help to address overoptimization and reduce the bias that relates to hyperparameter tuning and selection of the algorithm (Bradshaw et al., 2023). Since the cross validation approach involves splitting data into training and validation sets, when evaluating multiple models and determining their best hyperparameter values, using the validation error to estimate generalization error often leads to overestimating the model's performance. Thus, it is imperative to maintain a distinct test set to prevent using it for training or adjusting model parameters. The model's accuracy on this test set provides a reliable estimation of its performance on new or unseen data.

  The Nested Cross Validation (NCV) method is comprised of an outer cross validation loop and an inner cross validation loop. In the outer loop, the dataset is divided into multiple folds, in this work, was 5 folds. In each iteration, it uses one

fold as a test set and the rest of the folds are used for training. The model is trained on the training folds and it is evaluated on the test fold. This process repeats in each fold, resulting in multiple estimates of model performance. The inner loop involves selecting a training and validation set determined by the outer loop. The training folds are further split into multiple sub-folds and the model is trained using various hyperparameters on these training sets. The best hyperparameters are then determined based on the performance of the trained models on the validation set in the inner loop. This process ensures that the model's performance is accurately estimated and hyperparameters are optimized without overfitting (Maleki et al., 2020).

- *Evaluation Metrics*
  The evaluation metrics are a very useful and essential step when we want to evaluate our model performance, such as in this work on binary classification task. Evaluation metrics, including accuracy, precision, recall, specificity, F1 score, and ROC-AUC are used to evaluate the machine learning prediction model's performance. The most common metrics that are usually used in binary classification tasks are accuracy, sensitivity, specificity, and precision (Rainio et al., 2024). In this study, we evaluated the models' performance using the accuracy for the whole experiments, then, for further analysis of the comparison from three different inputs, we evaluated using the F1-score and ROC-AUC score. F1-score is the harmonic mean of Precision and Recall. We chose this metric because it is particularly useful when dealing with imbalanced dataset, as it provides a balance between precision and recall. Moreover, we chose the ROC-AUC score because it is also useful for dealing with imbalanced classes, as it compares the performance of classifiers, where it evaluates the True Positive Rate (TPR) and False Positive Rate (FPR).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Sensitivity (Recall/TPR)} = \frac{TP}{TP + FN} \tag{3}$$

$$F1\text{-}score = 2 \times \frac{precision \times recall}{precision + recall} \tag{4}$$

$$\text{FPR} = 1 - TPR = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{5}$$

*TP* True Positive, *TN* True Negative, *FP* False Positive, *FN* False Negative, *TPR* True Positive Rate, *FPR* False Positive Rate

## 4. Results

In this section, we present the results obtained from the classification and prediction task among dementia and MCI cases and the progression of the cognitive decline from MCI to dementia. The radiomics feature extraction computed 92 features from 7 feature classes in four image types (1 original image and 3 images applied filter) and a combination from all. Table 4 summarizes the number of features extracted from radiomics, applying preprocessing to clean the data before performing the feature extraction, and applying five different feature selection techniques.

We set up parameters for each technique, and for the PCA, we specified to select the number of principal components such that the total variance explained is at least 90% in the original data. In the Feature Importance using the random forest method, we selected the *n estimator* equal to 10, meaning we want to select the 10 most important features. While ANOVA and Mutual Information, we initialized *SelectKBest* with ANOVA F-value scoring function and Mutual Information Scoring for ANOVA and Mutual Information, respectively, to select the top k features, by default, the parameter $k$ is set to 10.

After the selected features were obtained, we trained our model with three different scenarios where we wanted to compare and evaluate the different inputs of the data to the model. In the first one, we ran the experiment using clinical variables alone, in the second one, we used radiomics features only, and the last one was a combination of both data. In total, we had 612 combinations to evaluate, in detail, we performed 12 different classifiers and 5 feature selection methods into 5 different inputs of image types in radiomics features alone and combined data (12 x 5 x 5 x 2 = 600 combinations). For the clinical variables, we used only the clinical data without applying any feature selection method (12 combinations). Tables 5, 6, 7, 8, and 9 show the results for different image types and the integrated data for each classifier and feature selection method.

From the results presented in the tables, we highlighted the highest accuracy in each image type, and as we can see, all the highest accuracy in each image type was consistency achieved by using the integrated data. Moreover, the best accuracy among all the experiments was achieved from the LoG image type using the KNN classifier with *kneighbors*: 15 and the *leaf size*: 10 and ANOVA feature selection method with an accuracy yielded of 0.83. Figure 4 describes the selected features using the ANOVA method based on ANOVA F-value scoring.

The features that were extracted are based on shape features (which in this study, were the most features extracted), intensity features, and volume/count features. Here are the explanations for each feature (van Griethuysen et al., 2017):

Table 4: The number of extracted features using radiomics and the selected features using different feature selection techniques

| Image type | Raw | Cleaned | PCA | FI | ANOVA | RFECV | MI |
|---|---|---|---|---|---|---|---|
| **Original** | 97 | 47 | 7 | 10 | 10 | 15 | 10 |
| **LoG** | 39 | 19 | 6 | 10 | 10 | 7 | 10 |
| **LBP3D** | 234 | 214 | 11 | 10 | 10 | 5 | 10 |
| **Wavelet** | 489 | 440 | 15 | 10 | 10 | 9 | 10 |
| **All images** | 819 | 735 | 18 | 10 | 10 | 4 | 10 |

*Cleaned*: Preprocessed data, *PCA*: Principal Component Analysis, *FI*: Feature Importance, *ANOVA*: Analysis of Variance, *RFECV*: Recursive Feature Elimination with cross validation, *MI*: Mutual Information

Table 5: Accuracy comparison of all classifiers for clinical variables and with five different feature selection methods using the original image type and the integrated data

| | Clinical var. dataset | | PCA | FI | ANOVA | RFECV | MI |
|---|---|---|---|---|---|---|---|
| **SVC** | 0.70 | Original Image | 0.73 | 0.73 | 0.75 | 0.75 | 0.66 |
| | | Integrated | 0.77 | 0.76 | 0.77 | 0.77 | 0.79 |
| **Linear SVC** | 0.66 | Original Image | 0.67 | 0.66 | 0.66 | 0.70 | 0.62 |
| | | Integrated | 0.71 | 0.71 | 0.70 | 0.70 | 0.69 |
| **RF** | 0.79 | Original Image | 0.73 | 0.74 | 0.72 | 0.76 | 0.69 |
| | | Integrated | 0.79 | 0.78 | 0.76 | 0.77 | 0.77 |
| **KNN** | 0.77 | Original Image | 0.77 | 0.75 | 0.74 | 0.75 | 0.73 |
| | | Integrated | 0.78 | **0.80** | 0.78 | 0.78 | 0.78 |
| **DT** | 0.74 | Original Image | 0.71 | 0.73 | 0.73 | 0.73 | 0.73 |
| | | Integrated | 0.70 | 0.72 | 0.72 | 0.73 | 0.73 |
| **LR** | 0.72 | Original Image | 0.72 | 0.72 | 0.72 | 0.72 | 0.72 |
| | | Integrated | 0.72 | 0.72 | 0.72 | 0.72 | 0.72 |
| **AB** | 0.74 | Original Image | 0.64 | 0.68 | 0.72 | 0.73 | 0.73 |
| | | Integrated | 0.71 | 0.75 | 0.75 | 0.75 | 0.75 |
| **GB** | 0.77 | Original Image | 0.73 | 0.73 | 0.73 | 0.77 | 0.73 |
| | | Integrated | 0.75 | 0.77 | 0.79 | 0.77 | 0.78 |
| **LDA** | 0.75 | Original Image | 0.67 | 0.72 | 0.74 | 0.73 | 0.73 |
| | | Integrated | 0.74 | 0.75 | 0.74 | 0.74 | 0.74 |
| **LDA shrinkage** | 0.78 | Original Image | 0.73 | 0.74 | 0.76 | 0.74 | 0.72 |
| | | Integrated | 0.75 | 0.75 | 0.77 | 0.78 | 0.74 |
| **QDA** | 0.75 | Original Image | 0.75 | 0.73 | 0.75 | 0.76 | 0.73 |
| | | Integrated | 0.74 | 0.75 | 0.77 | 0.74 | 0.75 |
| **GNB** | 0.65 | Original Image | 0.69 | 0.68 | 0.75 | 0.64 | 0.64 |
| | | Integrated | 0.68 | 0.67 | 0.73 | 0.68 | 0.69 |

*SVC*: Support Vector Classifier *RF*: Random Forest, *DT*: Decision Tree, *LR*: Logistic Regression, *AB*: Ada Boost, *GB*: Gradient Boosting, *LDA*: Linear Discriminant Analysis, *QDA*: Quadratic Discriminant Analysis, *GNB*: Gaussian Naïve Bayes, *Integrated*: Combined data (clinical variables and radiomics features)

- *original_shape_Sphericity*:
  The sphericity of an area refers to the degree of roundness in its shape. It computes the ratio between the volume of the shape and the volume of a sphere that has an equivalent surface area. This metric quantifies the roundness of the object, values approaching 1 mean a higher resemblance to a sphere.

- *original_shape_SurfaceVolumeRatio*:
  This metric measures the ratio of surface area to volume. Higher values indicate that the object has a greater surface area relative to its size, meaning a higher level of complexity or irregularity.

- *original_shape_MinorAxisLength*:
  The minimum length of the region's axis. It is the minimum distance spanning the object.

- *original_shape_Elongation*:
  It computes the ratio of the length axis between major and minor. This gives information about how much elongation of the object extent, larger numbers indicate a greater degree of stretching.

- *diagnostics_Image-original_Maximum*:
  It captures the highest intensity value that appears in the original image, which tells the brightest area in the image.

Table 6: Accuracy comparison of twelve classifiers for clinical variables and with five different feature selection techniques using the lbp3d image type and the integrated

| | Clinical var. dataset | | PCA | FI | ANOVA | RFECV | MI |
|---|---|---|---|---|---|---|---|
| **SVC** | 0.70 | LBP3D Image | 0.75 | 0.74 | 0.74 | 0.75 | 0.73 |
| | | Integrated | 0.77 | **0.81** | 0.78 | 0.77 | 0.74 |
| **Linear SVC** | 0.66 | LBP3D Image | 0.67 | 0.69 | 0.68 | 0.66 | 0.66 |
| | | Integrated | 0.70 | 0.74 | 0.70 | 0.71 | 0.66 |
| **RF** | 0.79 | LBP3D Image | 0.74 | 0.72 | 0.73 | 0.74 | 0.73 |
| | | Integrated | 0.75 | 0.76 | 0.76 | 0.78 | 0.75 |
| **KNN** | 0.77 | LBP3D Image | 0.73 | 0.75 | 0.75 | 0.75 | 0.75 |
| | | Integrated | 0.79 | 0.79 | 0.76 | 0.76 | 0.79 |
| **DT** | 0.74 | LBP3D Image | 0.72 | 0.73 | 0.70 | 0.73 | 0.70 |
| | | Integrated | 0.73 | 0.70 | 0.78 | 0.69 | 0.73 |
| **LR** | 0.72 | LBP3D Image | 0.72 | 0.72 | 0.72 | 0.72 | 0.72 |
| | | Integrated | 0.72 | 0.74 | 0.72 | 0.72 | 0.72 |
| **AB** | 0.74 | LBP3D Image | 0.73 | 0.71 | 0.71 | 0.73 | 0.74 |
| | | Integrated | 0.75 | 0.78 | 0.74 | 0.74 | 0.75 |
| **GB** | 0.77 | LBP3D Image | 0.73 | 0.74 | 0.74 | 0.76 | 0.72 |
| | | Integrated | 0.77 | 0.77 | 0.79 | 0.75 | 0.75 |
| **LDA** | 0.75 | LBP3D Image | 0.72 | 0.73 | 0.74 | 0.73 | 0.74 |
| | | Integrated | 0.74 | 0.77 | 0.76 | 0.77 | 0.73 |
| **LDA shrinkage** | 0.78 | LBP3D Image | 0.74 | 0.74 | 0.75 | 0.73 | 0.74 |
| | | Integrated | 0.75 | 0.79 | 0.75 | 0.76 | 0.74 |
| **QDA** | 0.75 | LBP3D Image | 0.71 | 0.74 | 0.74 | 0.71 | 0.75 |
| | | Integrated | 0.74 | 0.76 | 0.74 | 0.71 | 0.72 |
| **GNB** | 0.65 | LBP3D Image | 0.71 | 0.69 | 0.65 | 0.63 | 0.62 |
| | | Integrated | 0.69 | 0.71 | 0.65 | 0.68 | 0.67 |

*SVC*: Support Vector Classifier *RF*: Random Forest, *DT*: Decision Tree, *LR*: Logistic Regression, *AB*: Ada Boost, *GB*: Gradient Boosting, *LDA*: Linear Discriminant Analysis, *QDA*: Quadratic Discriminant Analysis, *GNB*: Gaussian Naïve Bayes, *Integrated*: Combined data (clinical variables and radiomics features)

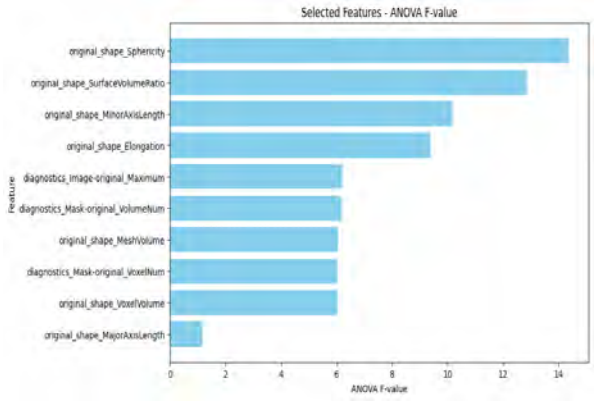



Figure 4: Selected features using ANOVA feature selection method based on ANOVA F-value scoring

- *diagnostics_Mask-original_VolumeNum*:
  The number that represents distinct volumes or regions identified within the mask of the original image.

- *original_shape_MeshVolume*:
  The volume of the region is determined by constructing a mesh around it. This provides the overall volume of the object in 3D space.

- *diagnostics_Mask-original_VoxelNum*:
  The overall voxel count, representing the number of 3D pixels, in the mask of the original image. This provides the overall number of 3D pixels making up the identified regions in the image.

- *original_shape_VoxelVolume*:
  The volume of the region is determined by multiplying the number of voxels by the size of each voxel. This provides information on the volume contained by the object in the 3D image, by quantifying the number of tiny cubes (voxels) that make it up.

- *original_shape_MajorAxisLength*:
  The maximum length of the region's axis. This measures the maximum distance across the object.

Additionally, to have better analysis and comparison, we ran our best model, which is the KNN as a classifier with the parameter *number of neighbors*: 15 and *leaf size*: 10, into our scenarios (three different inputs). Figure 5 illustrates the comparison of the model performance in each fold using the nested cross validation.

Table 7: Accuracy comparison of twelve classifiers for clinical variables and with five different feature selection techniques using the LoG image type and the integrated

| | Clinical var. dataset | | PCA | FI | ANOVA | RFECV | MI |
|---|---|---|---|---|---|---|---|
| **SVC** | 0.70 | LoG Image | 0.74 | 0.68 | 0.72 | 0.75 | 0.69 |
| | | Integrated | 0.77 | 0.75 | 0.77 | 0.81 | 0.81 |
| **Linear SVC** | 0.66 | LoG Image | 0.66 | 0.69 | 0.66 | 0.66 | 0.67 |
| | | Integrated | 0.71 | 0.72 | 0.71 | 0.71 | 0.71 |
| **RF** | 0.79 | LoG Image | 0.75 | 0.75 | 0.74 | 0.74 | 0.74 |
| | | Integrated | 0.77 | 0.77 | 0.81 | 0.79 | 0.79 |
| **KNN** | 0.77 | LoG Image | 0.74 | 0.75 | 0.75 | 0.75 | 0.75 |
| | | Integrated | 0.78 | 0.80 | **0.83** | 0.79 | 0.79 |
| **DT** | 0.74 | LoG Image | 0.73 | 0.73 | 0.73 | 0.73 | 0.73 |
| | | Integrated | 0.73 | 0.72 | 0.72 | 0.73 | 0.69 |
| **LR** | 0.72 | LoG Image | 0.72 | 0.72 | 0.72 | 0.72 | 0.72 |
| | | Integrated | 0.72 | 0.73 | 0.72 | 0.73 | 0.71 |
| **AB** | 0.74 | LoG Image | 0.70 | 0.72 | 0.73 | 0.73 | 0.71 |
| | | Integrated | 0.78 | 0.78 | 0.77 | 0.79 | 0.75 |
| **GB** | 0.77 | LoG Image | 0.76 | 0.73 | 0.73 | 0.76 | 0.74 |
| | | Integrated | 0.73 | 0.81 | 0.75 | 0.80 | 0.76 |
| **LDA** | 0.75 | LoG Image | 0.71 | 0.73 | 0.73 | 0.73 | 0.75 |
| | | Integrated | 0.75 | 0.75 | 0.74 | 0.77 | 0.76 |
| **LDA shrinkage** | 0.78 | LoG Image | 0.72 | 0.76 | 0.73 | 0.73 | 0.74 |
| | | Integrated | 0.77 | 0.77 | 0.77 | 0.77 | 0.75 |
| **QDA** | 0.75 | LoG Image | 0.70 | 0.74 | 0.73 | 0.71 | 0.74 |
| | | Integrated | 0.76 | 0.75 | 0.75 | 0.75 | 0.74 |
| **GNB** | 0.65 | LoG Image | 0.60 | 0.69 | 0.68 | 0.63 | 0.31 |
| | | Integrated | 0.66 | 0.72 | 0.71 | 0.36 | 0.71 |

*SVC*: Support Vector Classifier *RF*: Random Forest, *DT*: Decision Tree, *LR*: Logistic Regression, *AB*: Ada Boost, *GB*: Gradient Boosting, *LDA*: Linear Discriminant Analysis, *QDA*: Quadratic Discriminant Analysis, *GNB*: Gaussian Naïve Bayes, *Integrated*: Combined data (clinical variables and radiomics features)

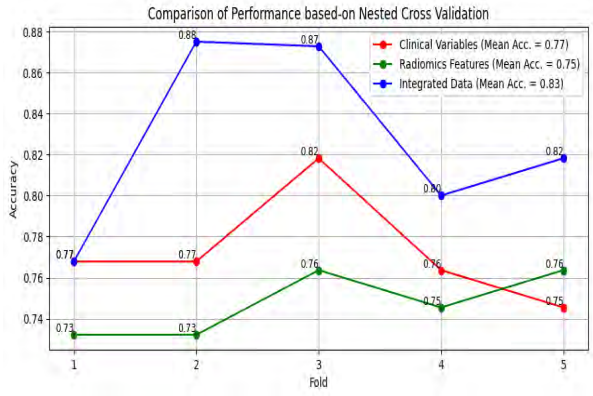



Figure 5: Comparison of performance in our scenarios, for the clinical variables with an average accuracy of 0.77, radiomics features with an average accuracy of 0.75, and the integrated of both data with an average accuracy of 0.83

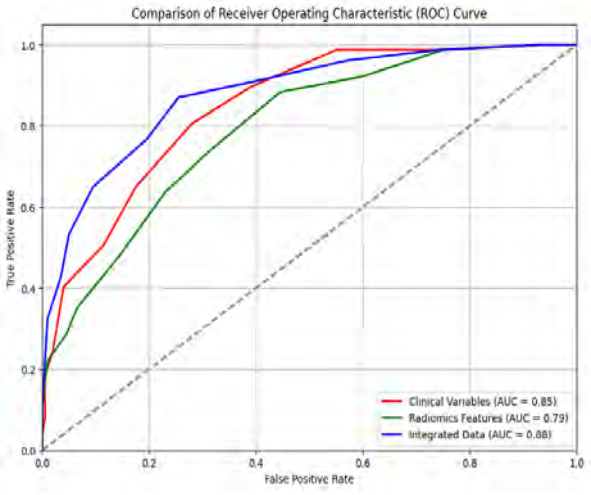



Figure 6: Comparison of model's performance in our scenarios using AUC curve, for the clinical variables with AUC score of 0.85, radiomics features with the AUC score of 0.79, and the integrated of both data yielded the among the scenarios with the AUC score of 0.88

Following that, we evaluated the effectiveness of our experiments by comparing the performance of their models with the ROC-AUC. Figure 6 illustrates the comparison of the model's performance using the ROC curve.

Based on the evaluation using the ROC-AUC, the integrated data still achieved the highest score compared to other scenarios. Additionally, Table 10 shows the comparison for all evaluation metrics we chose and the

Table 8: Accuracy comparison of twelve classifiers for clinical variables and with five different feature selection techniques using the wavelet image type and the integrated

| | Clinical var. dataset | | PCA | FI | ANOVA | RFECV | MI |
|---|---|---|---|---|---|---|---|
| **SVC** | 0.70 | Wavelet Image | 0.74 | 0.73 | 0.73 | 0.70 | 0.73 |
| | | Integrated | 0.76 | 0.75 | 0.79 | 0.79 | 0.78 |
| **Linear SVC** | 0.66 | Wavelet Image | 0.61 | 0.64 | 0.61 | 0.62 | 0.57 |
| | | Integrated | 0.70 | 0.72 | 0.73 | 0.71 | 0.66 |
| **RF** | 0.79 | Wavelet Image | 0.71 | 0.72 | 0.73 | 0.71 | 0.68 |
| | | Integrated | 0.79 | 0.78 | 0.78 | 0.78 | 0.77 |
| **KNN** | 0.77 | Wavelet Image | 0.75 | 0.74 | 0.73 | 0.75 | 0.72 |
| | | Integrated | 0.78 | 0.79 | **0.81** | 0.78 | 0.79 |
| **DT** | 0.74 | Wavelet Image | 0.70 | 0.75 | 0.68 | 0.73 | 0.71 |
| | | Integrated | 0.71 | 0.72 | 0.67 | 0.71 | 0.74 |
| **LR** | 0.72 | Wavelet Image | 0.72 | 0.72 | 0.72 | 0.72 | 0.72 |
| | | Integrated | 0.72 | 0.73 | 0.72 | 0.72 | 0.72 |
| **AB** | 0.74 | Wavelet Image | 0.71 | 0.69 | 0.73 | 0.72 | 0.71 |
| | | Integrated | 0.74 | 0.74 | 0.75 | 0.73 | 0.75 |
| **GB** | 0.77 | Wavelet Image | 0.72 | 0.73 | 0.74 | 0.72 | 0.75 |
| | | Integrated | 0.77 | 0.75 | 0.78 | 0.77 | 0.74 |
| **LDA** | 0.75 | Wavelet Image | 0.72 | 0.73 | 0.74 | 0.73 | 0.75 |
| | | Integrated | 0.75 | 0.75 | 0.73 | 0.75 | 0.76 |
| **LDA shrinkage** | 0.78 | Wavelet Image | 0.73 | 0.75 | 0.74 | 0.73 | 0.73 |
| | | Integrated | 0.77 | 0.76 | 0.74 | 0.75 | 0.77 |
| **QDA** | 0.75 | Wavelet Image | 0.71 | 0.72 | 0.74 | 0.73 | 0.75 |
| | | Integrated | 0.78 | 0.77 | 0.76 | 0.74 | 0.73 |
| **GNB** | 0.65 | Wavelet Image | 0.66 | 0.68 | 0.64 | 0.67 | 0.63 |
| | | Integrated | 0.70 | 0.68 | 0.67 | 0.69 | 0.68 |

*SVC*: Support Vector Classifier *RF*: Random Forest, *DT*: Decision Tree, *LR*: Logistic Regression, *AB*: Ada Boost, *GB*: Gradient Boosting, *LDA*: Linear Discriminant Analysis, *QDA*: Quadratic Discriminant Analysis, *GNB*: Gaussian Naïve Bayes, *Integrated*: Combined data (clinical variables and radiomics features)

AUC score. Lastly, Figure 7 reports the values of the confusion matrix from the best model.

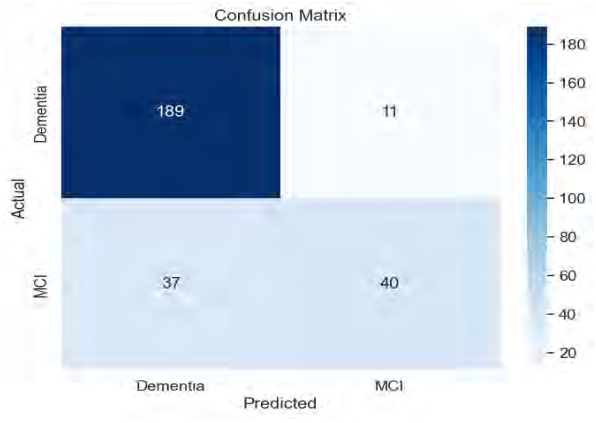



Figure 7: Heatmap of confusion matrix to show where the model is performing well and where it is not

The confusion matrix showed that the model performs well in predicting Dementia, with a high level of accuracy and precision, where the model correctly identified 189 out of 200 actual Dementia cases (True Positives), resulting in a precision of approximately 94%. The F1-score, which balances precision and recall, is approximately 89%, highlighting accurate prediction of dementia. However, the model misclassified 11 actual MCI cases as dementia (False Positives), which indicates that we can improve the performance to distinguish between the two classes. Overall, the model demonstrates reliable performance for identifying dementia cases by high precision and recall, although improvement can be made to reduce the false positive rate.

## 5. Discussion

Predicting the progression of cognitive decline is highly essential and helpful to understand the characteristics of the diseases and better strategy patient care. The early diagnosis might help to slow down the conversion of the MCI to dementia. Although several studies have explored different innovative methods of doing so, there are still a lot of works that need to be investigated, especially, exploring the analysis of utilizing $^{18}$F-FDG PET using radiomics feature extraction to predict cognitive decline. In recent times, Radimomics feature extraction has been successfully and widely used in medical imaging to extract a large amount of quantitative

Table 9: Accuracy comparison of twelve classifiers for clinical variables and with five different feature selection techniques using all image type and the integrated

| | Clinical var. dataset | | PCA | FI | ANOVA | RFECV | MI |
|---|---|---|---|---|---|---|---|
| **SVC** | 0.70 | All Image | 0.75 | 0.74 | 0.73 | 0.70 | 0.71 |
| | | Integrated | 0.77 | 0.79 | 0.79 | 0.73 | 0.73 |
| **Linear SVC** | 0.66 | All Image | 0.65 | 0.65 | 0.67 | 0.64 | 0.71 |
| | | Integrated | 0.69 | 0.74 | 0.71 | 0.69 | 0.67 |
| **RF** | 0.79 | All Image | 0.74 | 0.72 | 0.73 | 0.76 | 0.68 |
| | | Integrated | 0.77 | 0.76 | 0.80 | **0.81** | 0.77 |
| **KNN** | 0.77 | All Image | 0.75 | 0.74 | 0.74 | 0.76 | 0.70 |
| | | Integrated | 0.79 | **0.81** | **0.81** | 0.77 | 0.77 |
| **DT** | 0.74 | All Image | 0.66 | 0.74 | 0.72 | 0.75 | 0.59 |
| | | Integrated | 0.73 | 0.70 | 0.72 | 0.73 | 0.66 |
| **LR** | 0.72 | All Image | 0.72 | 0.72 | 0.72 | 0.72 | 0.68 |
| | | Integrated | 0.72 | 0.72 | 0.70 | 0.72 | 0.72 |
| **AB** | 0.74 | All Image | 0.71 | 0.73 | 0.70 | 0.73 | 0.68 |
| | | Integrated | 0.77 | 0.76 | 0.75 | 0.78 | 0.78 |
| **GB** | 0.77 | All Image | 0.75 | 0.76 | 0.71 | 0.74 | 0.69 |
| | | Integrated | 0.74 | 0.79 | 0.75 | 0.77 | 0.78 |
| **LDA** | 0.75 | All Image | 0.74 | 0.71 | 0.73 | 0.75 | 0.71 |
| | | Integrated | 0.76 | 0.77 | 0.75 | 0.75 | 0.74 |
| **LDA shrinkage** | 0.78 | All Image | 0.75 | 0.73 | 0.75 | 0.74 | 0.72 |
| | | Integrated | 0.75 | 0.77 | 0.75 | 0.76 | 0.75 |
| **QDA** | 0.75 | All Image | 0.69 | 0.69 | 0.75 | 0.77 | 0.68 |
| | | Integrated | 0.74 | 0.77 | 0.75 | 0.71 | 0.71 |
| **GNB** | 0.65 | All Image | 0.68 | 0.65 | 0.66 | 0.64 | 0.65 |
| | | Integrated | 0.71 | 0.69 | 0.70 | 0.68 | 0.76 |

*SVC*: Support Vector Classifier *RF*: Random Forest, *DT*: Decision Tree, *LR*: Logistic Regression, *AB*: Ada Boost, *GB*: Gradient Boosting, *LDA*: Linear Discriminant Analysis, *QDA*: Quadratic Discriminant Analysis, *GNB*: Gaussian Naïve Bayes, *Integrated*: Combined data (clinical variables and radiomics features)

Table 10: Comparison for all evaluation metrics we chose and the ROC-AUC score

| Scenario (input) | Metrics | | |
|---|---|---|---|
| | **Acc.** | **F1-score** | **AUC** |
| **Clinical variables** | 0.77 | 0.86 | 0.85 |
| **Radiomics features** | 0.75 | 0.84 | 0.79 |
| **Integrated data** | **0.83** | **0.89** | **0.88** |

information from the images. Many of the research was mainly focused on predicting the progression from MCI to AD using MRI brain images. In this study, we investigated and evaluated $^{18}$F-FDG PET brain images to predict not only the progression to AD but also to other subtypes of dementia that we pooled in one group of dementia. By extracting features using Pyradimocs framework, we have successfully extracted numerous amount of quantitative features that the radiologist could not see by looking at the $^{18}$F-FDG PET scans.

In this study, we proposed a two-step hierarchical binary classification. This approach served as a time-series classification model to predict the progression. We explored the use of machine learning classification models, which have a total of 12 different classifiers. To help find the most relevant features, we utilized five different feature selection methods. We aimed to analyze the results by comparing them with different inputs using our proposed method, which integrated clinical variables and radiomics features. We evaluated our model performance using nested cross-validation and evaluation metrics, including accuracy, F1-score, and AUC score. According to the results we obtained and showed above, our best model achieved 0.83, 0.89, and 0.88 for accuracy, F-1 score and AUC score, respectively. We achieved these results by using integrated data from LoG image type and the KNN as classifiers, with selected features from the ANOVA feature selection method. The experiments emphasize the overall model effectiveness by choosing suitable classifiers and feature selection methods. During the experiments, for the missing values in clinical variables, the KNN imputer resulted in a better outcome compared to the MICE technique. Moreover, regarding handling the imbalanced data, the results from the experiment showed that the model without applying the SMOTE and ADASYN techniques was better.

For future work, we can further analyze the use of $^{18}$F-FDG PET using radiomics features in a specific re-

gion of the brain. It might be very useful and more robust to investigate only in a specific region rather than the whole brain. Additionally, there is a study that uses white matter to extract the features. Investigating the extraction of features from each brain tissue may yield different and better results. Moreover, in terms of the radiomics, we can experiments and analyze of choosing the filters that we want to apply to the original image, also choosing the feature classes. Thus, we can investigate on the features extracted from the parameters that we define.

## 6. Conclusions

After doing several experiments, we can conclude that our study of predicting the enhancement of cognitive decline across three different scenarios - employing just clinical variables, radiomics features alone, and a combination of both has provided us with significant knowledge. At first, evaluating the clinical variables provided a fundamental understanding, offering insights into the traditional biomarkers or clinical variables of cognitive decline. Following that, the analysis of radiomics features allowed us to explore further into the complex details of $^{18}$F-FDG PET imaging data, potentially uncovering deeper patterns that indicate cognitive conversion. Lastly, the integration of clinical variables with radiomics features resulted in the most promising approach. Our investigations showed that by integrating clinical and radiomics data, we could enhance the accuracy and robustness of predictions regarding the progression of cognitive decline. The integration of this technique has substantial promise to enhance patient treatment and further our comprehension of neurodegenerative diseases.

## Acknowledgments

I would like to express my deepest gratitude to the European Union and the MAIA master's programme team for their helpful funding and unwavering support during the completion of a master's degree. Furthermore, I would like to extend a special thank you to my supervisors, Marco Bucci, Marina Bluma, and Caroline Dartora, for their tireless guidance, support, and valuable knowledge throughout the development of my master's thesis. Their support and guidance were critical in ensuring the successful completion of this research project. Lastly, I would like to extend my high appreciation to the Nordberg Translation Molecular Imaging Lab at Karolinska Institutet, particularly Prof. Agneta Nordberg for providing and allowing me to work with distinguished scholars in a prominent workplace.

## References

Alongi, P., Laudicella, R., Panasiti, F., Stefano, A., Comelli, A., Giaccone, P., Arnone, A., Minutoli, F., Quartuccio, N., Cupidi, C., Arnone, G., Piccoli, T., Grimaldi, L.M.E., Baldari, S., Russo, G., 2022. Radiomics analysis of brain [18f]fdg pet/ct to predict alzheimer's disease in patients with amyloid pet positivity: A preliminary report on the application of spm cortical segmentation, pyradiomics and machine-learning analysis. Diagnostics 12. URL: https://www.mdpi.com/2075-4418/12/4/933, doi:10.3390/diagnostics12040933.

Artaechevarria, X., Munoz-Barrutia, A., Ortiz-de Solorzano, C., 2009. Combination strategies in multi-atlas image segmentation: application to brain mr data. IEEE Transactions on Medical Imaging 28, 1266–1277. doi:10.1109/TMI.2009.2013857.

Ashraf, M.A., Goyal, A., 2023. Fludeoxyglucose (18F). StatPearls Publishing, Treasure Island (FL). URL: http://europepmc.org/books/NBK557653.

Awad, M., Fraihat, S., 2023. Recursive feature elimination with cross-validation with decision tree: Feature selection method for machine learning-based intrusion detection systems. Journal of Sensor and Actuator Networks 12, 67. doi:10.3390/jsan12050067.

Beraha, M., Metelli, A.M., Papini, M., Tirinzoni, A., Restelli, M., 2019. Feature selection via mutual information: New theoretical insights, in: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–9. doi:10.1109/IJCNN.2019.8852410.

Bevilacqua, R., Barbarossa, F., Fantechi, L., Fornarelli, D., Paci, E., Bolognini, S., Giammarchi, C., Lattanzio, F., Paciaroni, L., Riccardi, G.R., Pelliccioni, G., Biscetti, L., Maranesi, E., 2023. Radiomics and artificial intelligence for the diagnosis and monitoring of alzheimer's disease: A systematic review of studies in the field. Journal of Clinical Medicine 12, 5432. doi:10.3390/jcm12165432.

Bradshaw, T.J., Huemann, Z., Hu, J., Rahmim, A., 2023. A guide to cross-validation for artificial intelligence in medical imaging. Radiology. Artificial intelligence 5. doi:10.1148/ryai.220232.

Castiglioni, I., Rundo, L., Codari, M., Di Leo, G., Salvatore, C., Interlenghi, M., Gallivanone, F., Cozzi, A., D'Amico, N.C., Sardanelli, F., 2021. Ai applications to medical images: From machine learning to deep learning. Physics in Medicine 83, 9–24. doi:10.1016/j.ejmp.2021.02.006.

Centers for Disease Control and Prevention, 2011. Cognitive impairment: A call for action, now! [Online; accessed June 3, 2024].

Centers for Disease Control and Prevention, 2018. Subjective cognitive decline (scd). [Online; accessed April 24, 2024].

Cerami, C., Della Rosa, P.A., Magnani, G., et al., 2014. Brain metabolic maps in mild cognitive impairment predict heterogeneity of progression to dementia. NeuroImage: Clinical 7, 187–194. doi:10.1016/j.nicl.2014.12.004.

Chaddad, A., Desrosiers, C., Niazi, T., 2018. Deep radiomic analysis of mri related to alzheimer's disease. IEEE Access 6, 58213–58221. URL: https://api.semanticscholar.org/CorpusID:53093771.

Chan, H., Samala, R., Hadjiiski, L., Zhou, C., 2020. Deep learning in medical image analysis. Advances in Experimental Medicine and Biology 1213. doi:10.1007/978-3-030-33128-3_1.

Cheplygina, V., de Bruijne, M., Pluim, J.P.W., 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.

Chouliaras, L., O'Brien, J.T., 2023. The use of neuroimaging techniques in the early and differential diagnosis of dementia. Molecular Psychiatry doi:10.17863/CAM.100016.

Corey-Bloom, J., 2003. The abc of alzheimer's disease: cognitive changes and their management in alzheimer's disease and related dementias. International Psychogeriatrics 15, 33–49. doi:10.1017/s1041610203008664.

Dave, A., Hansen, N., Downey, R., Johnson, C., 2020. Fdg-pet imaging of dementia and neurodegenerative disease. Seminars in Ultrasound, CT and MRI 41, 49–57. URL: https://doi.org/10.1053/j.sult.2020.08.010, doi:10.1053/j.sult.2020.08.010.

Feng, F., Wang, P., Zhao, K., et al., 2018. Radiomic features of hippocampal subregions in alzheimer's disease and amnestic mild cognitive impairment. Frontiers in Aging Neuroscience 10, 290. doi:10.3389/fnagi.2018.00290.

Feng, Q., Niu, J., Wang, L., et al., 2021. Comprehensive classification models based on amygdala radiomic features for alzheimer's disease and mild cognitive impairment. Brain Imaging and Behavior 15, 2377–2386. doi:10.1007/s11682-020-00434-z.

Guedj, E., Varrone, A., Boellaard, R., Albert, N.L., Barthel, H., van Berckel, B., Brendel, M., Cecchin, D., Ekmekcioglu, O., Garibotto, V., Lammertsma, A.A., Law, I., Penuelas, I., Semah, F., Traub-Weidinger, T., van de Giessen, E., Van Weehaeghe, D., Morbelli, S., 2022. Correction to:eanm procedure guidelines for brain pet imaging using [f-18]fdg, version 3. European Journal of Nuclear Medicine and Molecular Imaging 49, 2100–2101. doi:10.1007/s00259-022-05755-3.

Jessen, F., Amariglio, R., Boxtel, M., Breteler, M., Ceccaldi, M., Chételat, G., Dubois, B., Dufouil, C., Ellis, K., Flier, W., Glodzik, L., Harten, A., de Leon, M., Mchugh, P., Mielke, M., Molinuevo, J., Mosconi, L., Osorio, R., Perrotin, A., Wagner, M., 2014. A conceptual framework for research on subjective cognitive decline in preclinical alzheimer's disease. Alzheimer's and dementia: the journal of the Alzheimer's Association 10. doi:10.1016/j.jalz.2014.01.001.

Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.W., 2010. elastix: A toolbox for intensity-based medical image registration. IEEE Transactions on Medical Imaging 29, 196–205. doi:10.1109/TMI.2009.2035616.

Koçak, B., Durmaz, E.Ş., Ateş, E., Kılıçkesmez, Ö., 2019. Radiomics with artificial intelligence: A practical guide for beginners. Diagnostic and Interventional Radiology 25, 485–495. doi:10.5152/dir.2019.19321.

Kumar, A., Sidhu, J., Goyal, A., Tsao, J.W., Doerr, C., 2023. Alzheimer Disease (Nursing). StatPearls Publishing, Treasure Island (FL). URL: http://europepmc.org/books/NBK568805.

Kumar, V., Gu, Y., Basu, S., Berglund, A., Eschrich, S.A., Schabath, M.B., Forster, K., Aerts, H.J., Dekker, A., Fenstermacher, D., Goldgof, D.B., Hall, L.O., Lambin, P., Balagurunathan, Y., Gatenby, R.A., Gillies, R.J., 2012. Radiomics: the process and the challenges. Magnetic Resonance Imaging 30, 1234–1248. doi:https://doi.org/10.1016/j.mri.2012.06.010. quantitative Imaging in Cancer.

Langa, K.M., Levine, D.A., 2014. The Diagnosis and Management of Mild Cognitive Impairment: A Clinical Review. JAMA 312, 2551–2561. doi:10.1001/jama.2014.13806.

Li, Y., Jiang, J., Lu, J., Jiang, J., Zhang, H., Zuo, C., 2019. Radiomics: a novel feature extraction method for brain neuron degeneration disease using 18f-fdg pet imaging and its implementation for alzheimer's disease and mild cognitive impairment. Therapeutic Advances in Neurological Disorders 12, 1756286419838682. doi:10.1177/1756286419838682.

Li, Y., Jiang, J., Shen, T., Wu, P., Zuo, C., 2018. Radiomics features as predictors to distinguish fast and slow progression of mild cognitive impairment to alzheimer's disease, in: Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 127–130. doi:10.1109/EMBC.2018.8512273.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical Image Analysis 42, 60–88. URL: http://dx.doi.org/10.1016/j.media.2017.07.005, doi:10.1016/j.media.2017.07.005.

Long, S., Benoist, C., Weidner, W., 2023. World Alzheimer Report 2023: Reducing Dementia Risk: Never Too Early, Never Too Late. Technical Report. Alzheimer's Disease International. URL: alzint.org/u/World-Alzheimer-Report-2023.pdf. [Online; accessed May 13, 2024].

Maleki, F., Muthukrishnan, N., Ovens, K., Reinhold, C., Forghani, R., 2020. Machine learning algorithm validation: From essentials to advanced applications and implications for regulatory certification and deployment. Neuroimaging Clinics of North America 30, 433–445. doi:https://doi.org/10.1016/j.nic.2020.08.004.

Mannil, M., von Spiczak, J., Manka, R., Alkadhi, H., 2018. Texture analysis and machine learning for detecting myocardial infarction in noncontrast low-dose computed tomography: Unveiling the invisible. Investigative Radiology 53, 338–343. doi:0.1097/RLI.0000000000000448.

Mayerhoefer, M.E., Materka, A., Langs, G., Häggström, I., Szczypiński, P., Gibbs, P., Cook, G., 2020. Introduction to radiomics. Journal of Nuclear Medicine 61, 488–495. URL: https://jnm.snmjournals.org/content/61/4/488, doi:10.2967/jnumed.118.222893.

Minoshima, S., Cross, D., Thientunyakit, T., Foster, N.L., Drzezga, A., 2022. 18f-fdg pet imaging in neurodegenerative dementing disorders: Insights into subtype classification, emerging disease categories, and mixed dementia with copathologies. Journal of Nuclear Medicine 63, 2S–12S. doi:10.2967/jnumed.121.263194.

Molnar, C., 2019. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Lulu.com. Available online: https://christophm.github.io/interpretable-ml-book/.

Pathan, M.S., Nag, A., Pathan, M.M., Dev, S., 2022. Analyzing the impact of feature selection on the accuracy of heart disease prediction. Journal of Medical Systems 48, 215–225. doi:10.1007/s10916-024-01567-9.

Peng, J., Wang, W., Song, Q., Hou, J., Jin, H., Qin, X., Yuan, Z., Wei, Y., Shu, Z., 2023. 18f-fdg-pet radiomics based on white matter predicts the progression of mild cognitive impairment to alzheimer disease: A machine learning study. Academic Radiology 30, 1874–1884. doi:https://doi.org/10.1016/j.acra.2022.12.033.

Perini, G., Rodriguez-Vieitez, E., Kadir, A., Sala, A., Savitcheva, I., Nordberg, A., 2021. Clinical impact of 18f-fdg-pet among memory clinic patients with uncertain diagnosis. European Journal of Nuclear Medicine and Molecular Imaging 48, 612–622. doi:10.1007/s00259-020-04969-7.

Rainio, O., Teuho, J., Klén, R., 2024. Evaluation metrics and statistical tests for machine learning. Scientific Reports 14, 6086. URL: https://doi.org/10.1038/s41598-024-56706-x, doi:10.1038/s41598-024-56706-x.

Rana, M., Bhushan, M., 2022. Machine learning and deep learning approach for medical image analysis: diagnosis to detection system. Medical Image Analysis 82. doi:10.1007/s11042-022-14305-w.

Rasi, R., A., G., Initiative, A.D.N., 2024. Predicting amyloid positivity from fdg-pet images using radiomics: A parsimonious model. Comput Methods Programs Biomed 247. URL: https://doi.org/10.1016/j.cmpb.2024.108098, doi:10.1016/j.cmpb.2024.108098.

Shu, Z.Y., Mao, D.W., yun Xu, Y., Shao, Y., Pang, P.P., Gong, X.Y., 2021. Prediction of the progression from mild cognitive impairment to alzheimer's disease using a radiomics-integrated model. Therapeutic Advances in Neurological Disorders 14, 17562864211029551. doi:10.1177/17562864211029551.

Singh, A., Kumar, R., Tiwari, A.K., 2023. Prediction of alzheimer's using random forest with radiomic features. Computer Systems Science and Engineering 45, 513–530. URL: http://www.techscience.com/csse/v45n1/49320, doi:10.32604/csse.2023.029608.

Sudharsan, M., Thailambal, G., 2023. Alzheimer's disease prediction using machine learning techniques and principal component analysis (pca). Journal of Medical Imaging and Health Informatics 14, 123–134. doi:10.1016/j.jmih.2024.01.005.

Teng, L., Li, Y., Zhao, Y., Hu, T., Zhang, Z., Yao, Z., Hu, B., (ADNI), A.D.N.I., 2020. Predicting mci progression with fdg-pet and cognitive scores: a longitudinal study. BMC Neurology 20, 148. URL: https://doi.org/10.1186/s12883-020-01728-x, doi:10.1186/s12883-020-01728-x.

Tixier, F., Vriens, D., Cheze-Le Rest, C., Hatt, M., Disselhorst, J.A., Oyen, W.J., de Geus-Oei, L.F., Visser, E.P., Visvikis, D., 2016. Comparison of tumor uptake heterogeneity characterization between static and parametric 18f-fdg pet images in non–small cell lung cancer. Journal of Nuclear Medicine 57, 1033–1039. doi:10.2967/jnumed.115.166918.

Ulaner, G., 2019. Fdg pet/ct performance and reporting,

in: Fundamentals of Oncologic PET/CT. Elsevier, pp. 5–8. doi:10.1016/b978-0-323-56869-2.00002-8.

van Griethuysen, J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R., Fillion-Robin, J.C., Pieper, S., Aerts, H., 2017. Computational radiomics system to decode the radiographic phenotype. Cancer Research 77, E104–E107. doi:10.1158/0008-5472.CAN-17-0339.

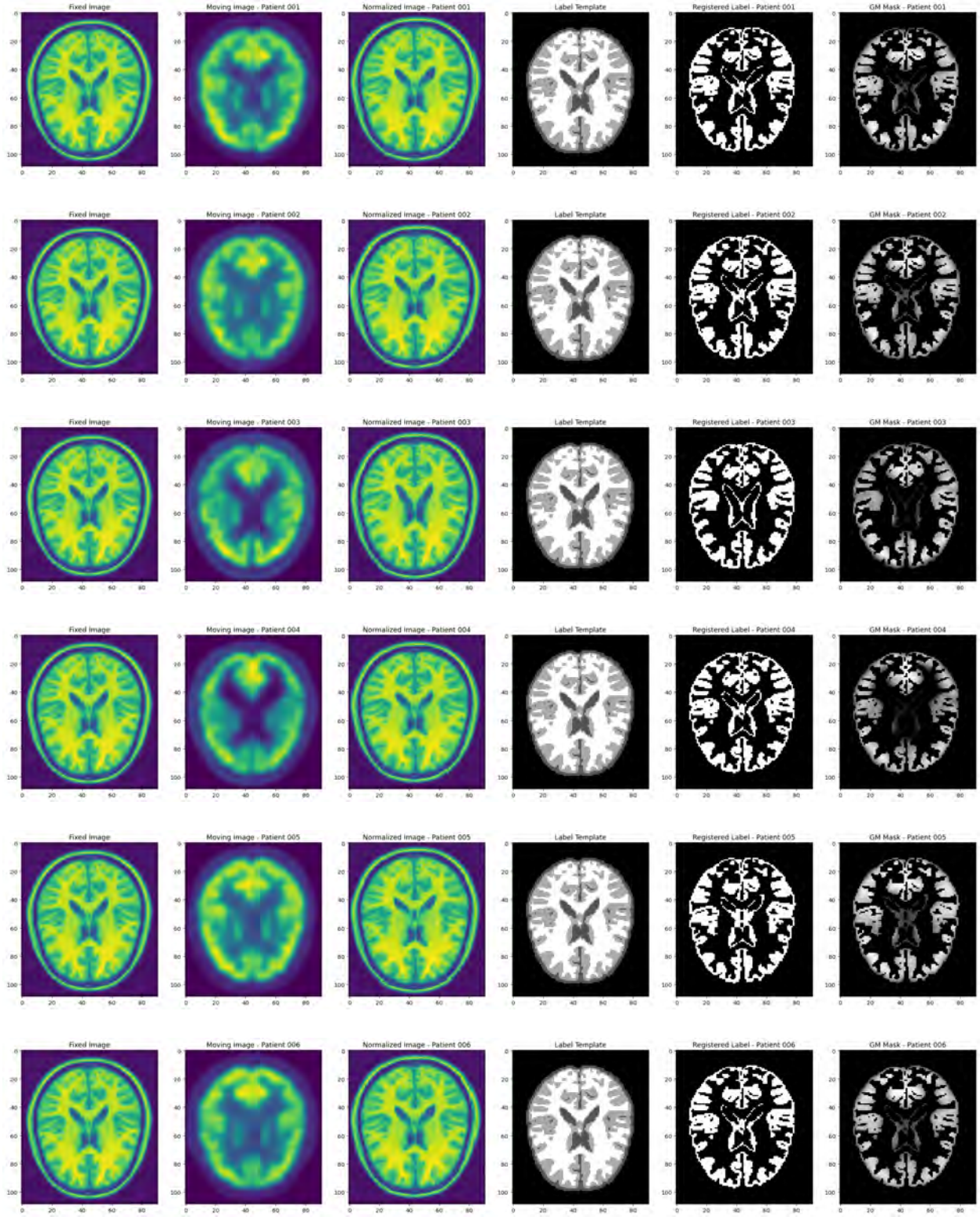
Varoquaux, G., Cheplygina, V., 2022. Machine learning for medical imaging: methodological failures and recommendations for the future. Nature Digital Medicine doi:10.1038/s41746-022-00592-y.

Vemuri, P., Lowe, V.J., Knopman, D.S., et al., 2016. Tau-pet uptake: Regional variation in average suvr and impact of amyloid deposition. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 6, 21–30. doi:10.1016/j.dadm.2016.12.010. published 2016 Dec 21.

Yakshit, Kaur, G., Kaur, V., Sharma, Y., Bansal, V., 2022. Analyzing various machine learning algorithms with smote and adasyn for image classification having imbalanced data, in: Proceedings of the 2022 International Conference on Computing, Communication, and Engineering Technologies (CCET), IEEE. pp. 1–7. doi:10.1109/CCET56606.2022.10080783.

## Appendix

### Examples of image processing in this work



**Parameter for Radiomics feature extraction**

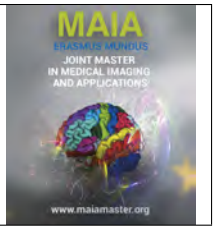
| Image type | Feature class | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Shape** | **First order** | **glcm** | **glrlm** | **glszm** | **ngtdm** | **gldm** |
| Original | MeshVolume | Mean | Autocorrelation | ShortRunEmphasis | SmallAreaEmphasis | Coarseness | GrayLevelNonUniformity |
| LBP3D | VoxelVolume | Median | ClusterProminence | LongRunEmphasis | LargeAreaEmphasis | Contrast | DependenceNonUniformity |
| LoG | SurfaceArea | StandardDeviation | ClusterShade | GrayLevelNonUniformity | GrayLevelNonUniformity | Busyness | GrayLevelVariance |
| Wavelet | SurfaceVolumeRatio | RootMeanSquared | ClusterTendency | RunPercentage | SizeZoneNonUniformity | Complexity | DependenceVariance |
| | Sphericity | RobustMeanAbsoluteDeviation | Contrast | GrayLevelVariance | ZonePercentage | Strength | DependenceEntropy |
| | Compactness1 | Variance | Correlation | RunVariance | GrayLevelVariance | | LowGrayLevelEmphasis |
| | Compactness2 | Uniformity | JointEnergy | RunEntropy | ZoneVariance | | HighGrayLevelEmphasis |
| | SphericalDisproportion | TotalEnergy | JointEntropy | LowGrayLevelRunEmphasis | ZoneEntropy | | |
| | Maximum3DDiameter | Entropy | Imc1 | HighGrayLevelRunEmphasis | LowGrayLevelZoneEmphasis | | |
| | Maximum2DDiameterSlice | | Imc2 | ShortRunLowGrayLevelEmphasis | HighGrayLevelZoneEmphasis | | |
| | Maximum2DDiameterColumn | | Idm | ShortRunHighGrayLevelEmphasis | SmallAreaLowGrayLevelEmphasis | | |
| | Maximum2DDiameterRow | | Idmn | LongRunLowGrayLevelEmphasis | SmallAreaHighGrayLevelEmphasis | | |
| | MajorAxisLength | | Id | LongRunHighGrayLevelEmphasis | LargeAreaLowGrayLevelEmphasis | | |
| | MinorAxisLength | | InverseVariance | | LargeAreaHighGrayLevelEmphasis | | |
| | LeastAxisLength | | MaximumProbability | | | | |
| | Elongation | | SumEntropy | | | | |
| | Flatness | | SumSquares | | | | |
| | MeshSurface | | | | | | |
| | PixelSurface | | | | | | |
| | Perimeter | | | | | | |
| | PerimeterSurfaceRatio | | | | | | |
| | Sphericity | | | | | | |
| | SphericalDisproportion | | | | | | |
| | MaximumDiameter | | | | | | |
| | MajorAxisLength | | | | | | |
| | MinorAxisLength | | | | | | |
| | Elongation | | | | | | |

**Parameter for affine registration:**

(FixedInternalImagePixelType ”float”)
(MovingInternalImagePixelType ”float”)
The dimensions of the fixed and moving image
(FixedImageDimension 3)
(MovingImageDimension 3)
(Registration ”MultiResolutionRegistration”)
(FixedImagePyramid ”FixedRecursiveImagePyramid”)
(MovingImagePyramid ”MovingRecursiveImagePyramid”)
(Interpolator ”BSplineInterpolator”)
(ResampleInterpolator ”FinalBSplineInterpolator”)
(Resampler ”DefaultResampler”)
(Optimizer ”RegularStepGradientDescent”)
(Transform ”AffineTransform”)
(Metric ”AdvancedMattesMutualInformation”)
(Scales 50000.0)
(AutomaticTransformInitialization ”false”)
The number of resolutions. 1 Is only enough if the expected deformations are small. 3 or 4 mostly works fine.
(NumberOfResolutions 3)
The pixel type of the resulting image
(ResultImagePixelType ”short”)
(ErodeMask ”false” ”false” ”false”)
(HowToCombineTransforms ”Compose”)
Number of spatial samples used to compute the mutual information in each resolution level.
(NumberOfSpatialSamples 30000 80000 100000)
(ImageSampler ”Random”)
Number of grey level bins in each resolution level, for the mutual information.
(NumberOfHistogramBins 16 32 32)
Order of B-Spline interpolation used in each resolution level:
(BSplineInterpolationOrder 2 2 2)
Order of B-Spline interpolation used for applying the final deformation.
(FinalBSplineInterpolationOrder 0)
Default pixel value for pixels that come from outside the picture:
(DefaultPixelValue 0)
Maximum number of iterations in each resolution level:
(MaximumNumberOfIterations 100 100 100)
Maximum step size of the RSGD optimizer for each resolution level.
(MaximumStepLength 4.0 2.0 2.0)
Minimum step size of the RSGD optimizer for each resolution level.
(MinimumStepLength 0.5 0.05 0.05)
Minimum magnitude of the gradient (stopping criterion) for the RSGD optimizer:
(MinimumGradientMagnitude 0.00000001 0.00000001 0.00000001)
Result image format
(ResultImageFormat ”nii”)

Medical Imaging and Applications

Master Thesis, June 2024




# Deep Spatiotemporal Models for the Assessment of Operative Difficulty in Laparoscopic Cholecystectomy Videos

Leonardo Pestana Legori[a], Saurav Sharma[a], Mario Scaglia[b], Maria Vannucci[c,d], Giovanni Guglielmo Laracca[e], Sergio Alfieri[f,g], Pietro Mascagni[d,f,g], Nicolas Padoy[a,d]

[a]*ICube, University of Strasbourg, CNRS, France*
[b]*Università degli Studi di Milano*
[c]*General Surgery Department, University of Torino, Turin, Italy*
[d]*IHU Strasbourg, Strasbourg, France*
[e]*Department of Medical Surgical Science and Translational Medicine, Sant'Andrea Hospital, Sapienza University of Rome, Rome, Italy*
[f]*Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy*
[g]*Università Cattolica del Sacro Cuore, Rome, Italy*

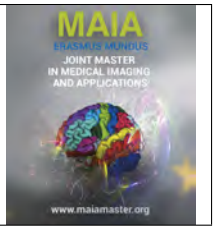
**Abstract**

Gallstone disease is a common diagnosis that can lead to life-threatening conditions if untreated. Laparoscopic cholecystectomy (LC) is the gold standard procedure for gallbladder removal, and, despite being safer than open surgery, major complications can still arise, leading to decreased patient survival and significant healthcare costs. The risks of complications are correlated with preoperative and intraoperative findings and, ultimately, the operative difficulty. Prediction of the LC operative difficulty (LCOD) could reduce the risk of adverse events by stratifying patients and assigning surgeons with the relevant skills. At the same time, there is a need to identify objective, clinically meaningful, and operator-independent definitions of LCOD. With that in mind, this study investigates deep spatiotemporal models for predicting LCOD in surgical videos, utilizing a novel dataset of 100 LC surgeries annotated with intraoperative features indicative of operative difficulty. We introduce spatiotemporal prediction pipelines that employ state-of-the-art deep learning architectures for both spatial and temporal sequence modeling. Our results demonstrate that spatiotemporal models enhance prediction performance compared to spatial-only models. These findings underscore the importance of temporal context in surgical video analysis and highlight the potential for improved intraoperative decision-making.

*Keywords:* Surgical workflow analysis, Laparoscopic surgery, Cholecystectomy, Deep learning, Transfer learning, Spatiotemporal modeling

Medical Imaging and Applications

Master Thesis, June 2024




# Impact of Lesion Inclusion on Biomechanical Modeling Using Deep Learning-based Breast Tissue Segmentation

Melika Pooyan, Eloy García Marcos, Robert Martí Marly

*University of Girona, VICOROB laboratory*

**Abstract**

The integration of 2D mammography with 3D Magnetic Resonance Imaging (MRI) is crucial for enhancing the accuracy of breast cancer diagnosis and treatment planning. However, this integration poses significant challenges due to the inherent differences in imaging modalities and the need for precise tissue segmentation and alignment. This thesis addresses these challenges by focusing on the geometry extraction and multi-class segmentation of breast tissues from MRI data. Utilizing the nnU-Net architecture, our study achieves segmentation performance for breast tissues comparable to state-of-the-art, with Dice coefficients of 0.94, 0.88, and 0.87 for fat, glandular tissue, and pectoral muscle, respectively. Notably, the overall foreground achieves a mean Dice coefficient of 0.83 using an ensemble of 2D and 3D U-Net configurations. These high dice coefficients ensure the accuracy required for subsequent 3D reconstruction and biomechanical modeling. The segmented data is then used to generate detailed 3D meshes and develop biomechanical models using NiftySim, which simulates breast tissue's physical properties and behaviors. Furthermore, the research explores the biomechanical behavior of these models in the presence of benign and malignant lesions, providing insights into how different tissue types interact under various conditions. The findings of this study have the potential to improve the integration of 2D and 3D imaging modalities such as X-ray and MRI, thereby enhancing diagnostic accuracy and treatment planning for breast cancer.

*Keywords:* Multi-class Tissue Segmentation, nnU-Net, Biomechanical Modeling, Lesion Characterization, Mechanical Deformations

## 1. Introduction

Breast cancer is the most common cancer among women, with 1 in 8 women developing invasive breast cancer in their lifetime, highlighting the need for early and accurate diagnosis to improve patient outcomes (Smith, 2013). While traditional imaging techniques provide valuable information, they have inherent limitations. Advanced methods like multi-modality imaging powered by image registration correspondence can address these limitations by fusing data from different sources, leading to a more comprehensive analysis.

Combining imaging techniques such as mammography and MRI provides a comprehensive view of the breast, improving diagnosis and treatment planning. Mammography detects microcalcifications but struggles with dense tissue, whereas MRI excels in soft tissue contrast and detecting invasive cancers. Integrating these modalities enhances lesion detection and characterization (García et al., 2018). However, differences in patient positioning during imaging, such as mammographic compression and prone positioning in MRI, present challenges (van Engeland et al., 2003; Pinto Pereira et al., 2010; Rueckert et al., 1999). Advanced image registration techniques are needed to align these images accurately (Arlinghaus et al., 2011; Siegler et al., 2012). Finite Element Analysis (FEA) helps simulate breast tissue deformation under different conditions, aiding in accurate image registration. Patient-specific models replicating the breast's physical properties improve the precision of diagnostic and therapeutic interventions (Babarenda Gamage et al., 2012; Garcia et al., 2017; Melbourne et al., 2011). Despite advancements, the deformable nature of breast tissue complicates image correlation across modalities and clini-

cal contexts, affecting diagnosis, biopsy guidance, and surgical planning (Garcia et al., 2017). Biomechanical modeling offers valuable insights into breast tissue behavior, understanding disease progression, and treatment planning. However, accurately identifying different tissue types within patient-specific models derived from 3D modalities like MRI is a time-consuming and error-prone manual task (García et al., 2018). Due to its high soft-tissue contrast, MRI can detect the discrimination between different structures in the breast and enable 3D visualization (Giess et al., 2014). However, breast MRI imaging includes other organs such as the lung, heart, pectoral muscles, and thorax. As a result, it is crucial to segment the breast region from the other organs to ensure accurate analysis in biomechanical modeling.

This work explores the potential of automating tissue segmentation using deep learning techniques, specifically the nnU-Net framework (Isensee et al., 2021). nnU-Net offers several advantages for this application. First, it automatically configures its architecture based on the specific dataset, eliminating the need for extensive manual tweaking. This allows for the efficient processing of diverse breast DCE-MRI datasets. Second, by employing nnU-Net, we aim to achieve higher accuracy in tissue segmentation, leading to improved mesh quality in the biomechanical models. Finally, Our proposed segmentation approach incorporates lesion information into existing biomechanical models, a capability not previously explored. This innovation significantly enhances the performance of existing models by explicitly accounting for lesions. Our hypothesis is to assign specific material properties to lesions for a more detailed biomechanical analysis of their impact on breast tissue behavior. In essence, this work helps to bridge the gap between advanced medical imaging techniques and clinical applications in breast cancer. By automating segmentation with deep learning and incorporating lesion data, we aim to develop more accurate and informative biomechanical models, potentially leading to improved outcomes for breast cancer patients Figure 1 shows the step's workflow.

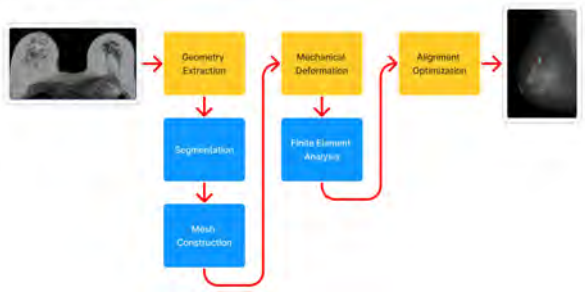



Figure 1: A general workflow of registering MRI to X-ray mammography images. The procedure encompasses geometry extraction, mechanical deformation, and final alignment with X-ray mammography. This thesis primarily aims to enhance the geometry extraction step. The work is based on the work of (García et al., 2018).

## 2. State of the art

### 2.1. *Multi-Modality Correspondence: A Challenge in MRI-Mammography Integration*

One of the critical advancements in breast cancer diagnosis is the integration of different imaging modalities to overcome limitations and enhance diagnosis. This approach, particularly the registration of Magnetic Resonance Imaging (MRI) with mammography, has emerged as a pivotal technique. MRI offers a more detailed view of breast tissue, revealing lesions and tissues in 3D, aiding tumor assessment (size, shape, spread) than mammograms (Giess et al., 2014). Combining these modalities offers a comprehensive view, enhancing diagnosis and treatment planning. This leverages the strengths of both techniques, improving breast cancer detection sensitivity and specificity (Garcia et al., 2017). However, achieving precise correspondence (alignment) between MRI and mammogram data presents a significant challenge. The core challenge lies in the inherent differences between the modalities. MRI captures detailed 3D structures, while mammography provides high-resolution 2D slices during compression, distorting the 3D geometry. This makes it difficult to map structures in one modality to the other precisely. Hence, patient-specific finite element (FE) models have emerged as a promising solution (García et al., 2018). These models can simulate the mammogram acquisition process by compressing a 3D MRI and projecting lesions onto a 2D plane. However, localizing the exact 3D location of the lesion in the MRI based solely on the mammogram remains challenging. Traditional registration algorithms often struggle with this reverse mapping (Solves-Llorens et al., 2014b). Recent advancements utilize both craniocaudal (front-to-back) and mediolateral oblique (angled) mammographic views to calculate the X-ray path to the lesion within the compressed breast model (García et al., 2017). This, combined with using barycentric coordinates within the model's elements, allows for more accurate 3D localization of the lesion in the MRI (García et al., 2017). This improved accuracy translates to essential information for guiding clinical decision-making (Solves-Llorens et al., 2014b). Despite these advancements, achieving perfect multi-modality correspondence remains challenging due to variations in patient positioning, compression levels, and image acquisition settings, all introducing errors. While current techniques can minimize these errors, ongoing research is crucial to further refine the process and ensure its wider clinical adoption (Solves-Llorens et al., 2014b). Ultimately, by overcoming the multi-modality correspondence challenge, we can unlock the full potential of combining MRI and mammography for enhanced breast cancer detection and treatment strategies.

### 2.2. *Segmentation Techniques of Breast MRI: Traditional Techniques*

Most studies, as summarized in Table 1, focus on segmenting specific parts of the breast, such as the pectoral muscle or glandular tissue, rather than addressing the whole breast region in raw MRI data. This targeted segmentation approach might overlook the interactions between different tissue types, which are crucial for comprehensive breast cancer diagnosis. The Dice coefficients reported in these studies vary, indicating differences in segmentation accuracy across different methodologies and imaging modalities. For example, methods like U-Net and nnU-Net achieve relatively high Dice coefficients, whereas traditional approaches such as Fuzzy C-Means and Expectation-Maximization show lower performance.

Traditional segmentation methods, such as atlas-based and template-based approaches, have been widely used to address the challenges in breast MRI segmentation. Atlas-based segmentation leverages pre-labeled anatomical atlases registered to a patient's MRI scan, useful for handling anatomical variability across patients. For example, in breast MRI, (Gubern-Mérida et al., 2012) developed an atlas-based method for segmenting the pectoral muscle. Fuzzy C-Means (FCM) clustering, which assigns each voxel to a class based on distance in the feature space, has also been applied in breast MRI segmentation. While these traditional methods have been effective, their performance is generally lower compared to modern deep-learning techniques.

#### 2.2.1. *Deep Learning-Based Segmentation Techniques*

Deep learning techniques have transformed medical image segmentation, offering high accuracy and efficiency. These methods use large datasets to train neural networks for automatic segmentation with minimal user intervention. The U-Net architecture, a convolutional neural network (CNN) designed for biomedical image segmentation, captures both high-level and fine-grained features. Studies such as those by (Zhang et al., 2019) and (Alqaoud et al., 2022a) have shown that U-Net and its variants can achieve significantly higher Dice coefficients for segmenting glandular tissue and other structures compared to traditional methods.

Recent advancements have focused on improving segmentation accuracy for specific breast tissues. For instance, transformer-based neural networks and enhanced architectures like nnU-Net have demonstrated superior performance. (Müller-Franzes et al., 2023) and (Alqaoud et al., 2022a) reported high accuracy in segmenting glandular tissue, showcasing the effectiveness of these advanced techniques. Furthermore, multi-class segmentation, including fat, glandular, and tumorous tissues, has been explored using deep neural networks, indicating promising results for comprehensive breast tissue analysis.

Table 1 offers a summary of these segmentation techniques, highlighting the shift from traditional to deep learning methods and the associated improvements in segmentation performance across various studies.

### 2.3. *Mesh Generation and Biomechanical Modeling*

While achieving accurate segmentation is the primary focus of this thesis, its ultimate goal is to leverage these results for building more realistic biomechanical breast models. This necessitates mesh generation as the next crucial step. The process begins with creating a surface mesh from the segmented MRI images. Techniques like B-splines are used to create smooth 2D contours that form a 3D surface mesh, while algorithms such as marching cubes generate a triangular mesh representing the breast surface (Chung et al., 2008; Zhang et al., 2007).

Next, a volumetric mesh defines the breast's internal structure using tetrahedral or hexahedral elements, selected based on the model's requirements for geometric flexibility or simulation stability (del Palomar et al., 2008; Solves-Llorens et al., 2014b). This step ensures that the internal anatomy of the breast is accurately captured, which is essential for realistic simulations.

Finite Element Analysis (FEA)is then used to simulate the mechanics of breast tissue. FEA solves equations that describe how the tissue deforms under various conditions, such as gravity and mammographic compression. Patient-specific FEA models are constructed from 3D MRI-based structures, segmenting breast tissues and incorporating mechanical properties like elasticity and stiffness. These models aid in image registration and enhance lesion localization, thereby improving diagnostic accuracy (García et al., 2018; Mertzanidou et al., 2011; Pathmanathan et al., 2008; Solves-Llorens et al., 2014a).

Current limitations in breast imaging technologies include the lack of large, standardized datasets containing both MRI and mammogram images from the same patients. This makes the development of robust models challenging, particularly for tasks like accurate segmentation and lesion characterization. Additionally, discrepancies in patient positioning between MRI and mammogram scans further complicate image registration, impacting the overall analysis. Furthermore, limitations exist within the modeling techniques themselves. FEA tools can be computationally expensive and require significant expertise for accurate mesh generation and material property assignment. Additionally, specific FEA software, like NiftySim, might be sensitive to anatomical features, potentially failing for breasts with smaller sizes due to nipple-to-chest wall distance limitations.

Overcoming these challenges requires advancements in two key areas: data acquisition strategies and modeling techniques. By focusing on acquiring standardized, well-aligned data from both modalities and developing

| Study | Segmented Classes | Methodology | Imaging Modalities | Dice Coefficient |
|---|---|---|---|---|
| Gubern-Mérida et al. (2012) | Pectoral muscle | Atlas-based | Breast MRI | 0.74 (multi-atlas) and 0.72 (probabilistic) |
| Zafari et al. (2019) | **Pectoral muscle** | U-Net | Breast MRI | **0.89** |
| Zhang et al. (2019) | Glandular Tissue | U-Net | Breast MRI | 0.83 ± 0.04 |
| Müller-Franzes et al. (2023) | Glandular Tissue | Transformer-based neural network | Breast MRI | 0.864 ± 0.081 |
| Alqaoud et al. (2022a) | **Glandular Tissue** | nnU-Net | Multi-modality MRI | **0.877 ± 0.081** |
| Huo et al. (2021) | Glandular Tissue | nnU-Net | DCE-MRI | 0.85 |
| Razavi et al. (2015) | Glandular Tissue | Fuzzy C-Means | Breast MRI | 0.84 |
| Gubern-Mérida et al. (2015) | Glandular Tissue | Expectation-maximization | Breast MRI | 0.80 |
| Wu et al. (2012a) | Glandular Tissue | Fuzzy C-Means | Breast MRI | 0.73 |
| Wu et al. (2012b) | Glandular Tissue | Atlas-aided Probabilistic | Breast MRI | 0.85 |
| Zhang et al. (2019) | Glandular Tissue | U-Net | Breast MRI | 0.83 ± 0.06 |
| Dalmış et al. (2017) | Glandular Tissue | 3C U-Net | Breast MRI | 0.85 |
| Alqaoud et al. (2022b) | **Fat**, Glandular Tissue, and tumorous tissue | DNN | Breast MRI | **0.95**, 0.83, and 0.41 |

Table 1: Summary of Traditional and Deep Learning Segmentation Techniques with Dice Coefficients

robust models that can effectively segment and explore lesions, we can significantly improve the precision and clinical utility of breast imaging technologies (García et al., 2018). This research aims to address these limitations by proposing methods that enhance segmentation accuracy, enabling more robust exploration of lesions, ultimately leading to improved breast cancer diagnosis and treatment.

## 3. Material and methods

### *3.1. **Dataset***

The dataset used for the study comprised 166 T1-weighted non-fat saturated dynamic contrast-enhanced MRI (DCE-MRI) scans from various patients, including follow-up scans. DCE-MRI is a type of MRI scan that uses a contrast agent to track blood flow and identify abnormalities. It involves both pre-contrast images and post-contrast images (taken after contrast injection) to see how tissues take up the contrast. The MRI scans were acquired using a dedicated 1.5 Tesla Siemens Magnetom Vision system with a CP Breast Array coil. While the scans varied slightly in terms of pixel spacing and slice thickness (ranging from 0.625 to 0.722 mm spacing, 1.3 mm slice thickness, and a volume of 512x256x120 voxels), they all followed a standardized protocol. Patients were positioned face-down for the scans. Each patient's DCE-MRI data was originally stored in a DICOM series. This series combined both pre-contrast and post-contrast images into a single volume with separate channels for each. We then separated these channels into individual pre-contrast and

post-contrast volumes using the SimpleITK library. Pre-contrast volumes are ideal for tissue segmentation because they offer a clearer visualization of tissue structure before the contrast agent is introduced. Conversely, the last post-contrast volume is preferred for lesion detection due to the enhanced visibility of lesions after contrast administration.

An experienced observer accurately segmented the 166 MRI volumes into seven distinct categories: background, fatty tissue, glandular tissue, heart, lung area, pectoral muscles, and thorax. Two additional observers focused solely on segmenting the pectoral muscles. The segmentation process involved manually labeling every 5-10 slices of an MRI volume. To create complete labeling, linear interpolation filled in the gaps between manually labeled slices. Structures requiring more precise definition, like the heart, lungs, and pectoral muscles, were segmented with a smaller slice interval during the manual labeling stage. Background, fatty, and glandular tissue segmentation employed thresholding techniques on regions chosen by the observer (Gubern-Mérida et al., 2012).

The dataset also included a subset of 10 cases specifically chosen to test the framework's ability to handle lesions. These case's ground truths are primarily diagnosed with benign lesions like fibroadenoma and water cysts. Additionally, four cases suspected of having malignant lesions were identified from this subset and annotated using ITK-SNAP software. A radiologist then validated the annotations for these four cases, confirming that two were benign and two were malignant lesions. Figure 2 illustrates an example from the dataset.

Figure 3 shows the distribution of different tissue types within the dataset. The bars represent the number of voxels (3D image units) categorized as each tissue type. We can observe that fat and thorax are the most dominant classes, while pectoral muscle and other tissue classes appear less frequent.

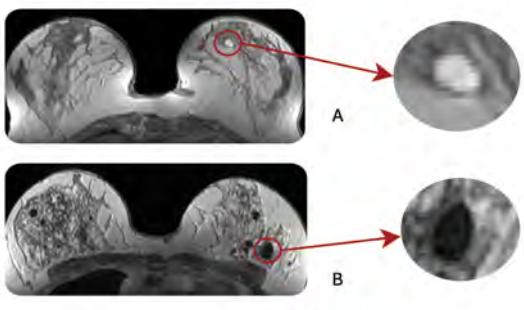



Figure 2: Breast MR scan on an axial slice with two cases of (A) malignant and (B) benign lesion.

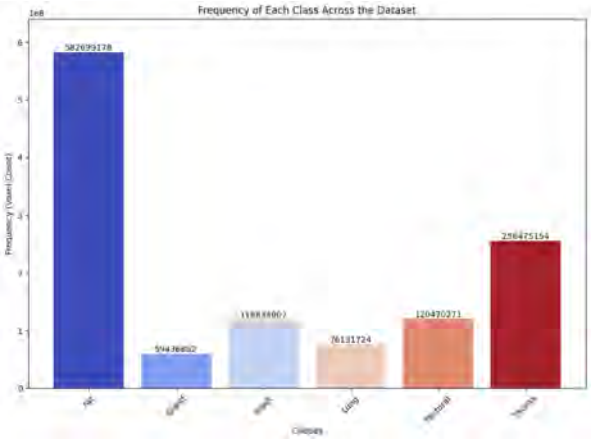



Figure 3: Frequency of each class in a dataset containing each class voxel counts.

### 3.2. *Methods*

#### 3.2.1. *Segmentation*

nnU-Net (no-new-UNet) has emerged as a leading method for medical image segmentation tasks (Isensee et al., 2021). It builds upon the success of the original U-Net architecture, known for its encoder-decoder design that effectively combines both spatial and semantic information through skip connections. nnU-Net offers a flexible toolbox of U-Net variations, including 2D U-Net, 3D U-Net, and a U-Net Cascade. While both 2D and 3D U-Nets directly generate full-resolution segmentation masks, the cascade approach first generates lower-resolution segmentations and then refines them for improved accuracy (Isensee et al., 2021). Our study leverages both 2D and 3D U-Net architectures for training with five-fold cross-validation with 100 epochs per fold. There was no need to use cascade since the 3D-UNet patches could capture the whole image. Here's a closer look at each approach:

- **2D U-Net:** This configuration shares a similar architecture with the original U-Net and it runs on full-resolution data and it is expected to work well on anisotropic data. For 3D datasets, we strategically extract 2D slices (typically from the plane with the highest resolution) and train the neural network on these individual slices.
- **3D U-Net** While a popular choice for 3D segmentation tasks running on full resolution data. 3D U-Net can be limited by graphics processing unit (GPU) memory constraints. When dealing with large datasets, we needed to segment the data into smaller 3D patches for training as input. This approach, however, can lead to a loss of valuable contextual information within the original 3D data.

nnU-Net utilizes a unique pipeline to prepare and train the segmentation models. This pipeline starts by analyzing the training data to create a "data fingerprint"

a set of characteristics specific to the data. Based on this fingerprint, the pipeline automatically selects appropriate hyper-parameters such as the loss function, optimizer, and network architecture (Isensee et al., 2021). nnU-Net uses a combination of cross-entropy loss and dice loss functions known as a compound loss function to train 6 classes, hence the segmentation accuracy and training stability will increase (Isensee et al., 2021). Moreover, nnU-Net utilizes the stochastic gradient descent method with initial learning (0.01) and Nesterov Momentum (0.9) to optimize the loss function (Isensee et al., 2021). Regarding pre-processing steps like image resampling, normalization, and the size of data batches and patches are used for training (Isensee et al., 2021). These choices, along with the data fingerprint, form a unique "pipeline fingerprint" Table 2 shows the dataset fingerprint created by the nnUNet and each configuration hyper-parameters. Leveraging this pipeline fingerprint, nnU-Net trains separate models using both 2D and 3D U-Net architectures. Each model is trained with the hyper-parameters determined earlier. Since there is a Class imbalance in the dataset (Figure 3) it is addressed by oversampling foreground regions while combining Dice loss with cross-entropy loss for improved training stability and accuracy (Isensee et al., 2021).To enhance performance, the final prediction is generated by combining (ensemble) the outputs from 2D and 3D U-Net network configurations which is done by averaging the softmax probabilities between the segmentation output of two configurations to generate the final segmentation labels. This ensemble is evaluated based on the Dice coefficient, a metric that measures segmentation accuracy, on the training data. Ultimately, the best-performing ensemble configuration is used to generate predictions for unseen data (test set). Figure 4 illustrates the nnU-Net pipeline used for this study, indicating the training process with 166 cases and a test set of 10 cases.

| **Dataset** | |
|---|---|
| Median image size | 120x254x510 |
| Median image spacing | 1.29x0.66x0.66mm |
| Normalization | Z-score |

| **2D-UNet** | |
|---|---|
| Target Spacing | NAx254x510 |
| Median Shape @ Target Spacing | NAx0.66x0.66mm |
| Patch Size | 256x512 |
| Batch Size | 24 |

| **3D-fullres UNet** | |
|---|---|
| Target Spacing | 120x254x510 |
| Median Shape @ Target Spacing | 1.29x0.66x0.66mm |
| Patch Size | 64x128x288 |
| Batch Size | 2 |

Table 2: Dataset fingerprint and configurations for 2D-UNet and 3D-fullres UNet.

#### *3.2.2.* ***Geometry Extraction and Mesh Generation***

The second step in this study aimed to incorporate lesions into the segmentation maps. While a deep learning approach like nnUnet is powerful for segmentation tasks, in this initial analysis, we opted for a simpler approach due to limitations in the available dataset. The dataset currently does not contain enough training data to achieve optimal performance for lesion segmentation using complex deep-learning architectures. However, lesion detection and segmentation are crucial aspects of our research, and we plan to address this limitation in future work. Building upon our earlier work on DCE-MRI-based lesion segmentation (Vidal et al., 2022), we aim to develop and integrate a robust deep-learning framework for accurate lesion segmentation within the biomechanical modeling pipeline.

Here's how we achieved our current approach:

1. We first used the trained nnU-Net model to obtain initial tissue segmentation for each case.
2. Next, we combined these initial tissue segmentations with the manually annotated lesion ground truths.
3. This process resulted in new segmentation maps that included both tissue types and lesions, allowing for further analysis.

This section draws inspiration from the work of (Garcia et al., 2017), regarding geometry extraction for biomechanical breast modeling. Accurate segmentation is crucial for this process. We begin by utilizing nnUNet segmentation results from the previous step, which identifies various anatomical structures like the pectoral muscle, lungs, heart, thorax, and breast tissue. Following the nnUNet application to MRI volumes to exclude non-breast tissues, we applied a breast region mask obtained from the region-growing algorithm to segment the image background, hence the breast became isolated with containing volumes of interest which are fat, and glandular tissue, with the sternum serving as a reference point as suggested in (Gubern-Mérida et al., 2012).

Following this, the isolated breast volume segmentation map with its internal fat and glandular tissues is resampled to isotropic voxels of $1\,mm^3$ for mesh efficiency. The volume mesh is generated using pygalmesh (Schlömer, 2021), a Python interface for CGAL's (The CGAL Project, 2024), which is capable of generating 2D and 3D meshes. The element count varies between 50,000 and 500,000 depending on breast volume (Garcia et al., 2017), minimizing errors during finite element simulations (Del Palomar et al., 2008). Figure 5 shows the result of the preprocessing of the mesh generation.

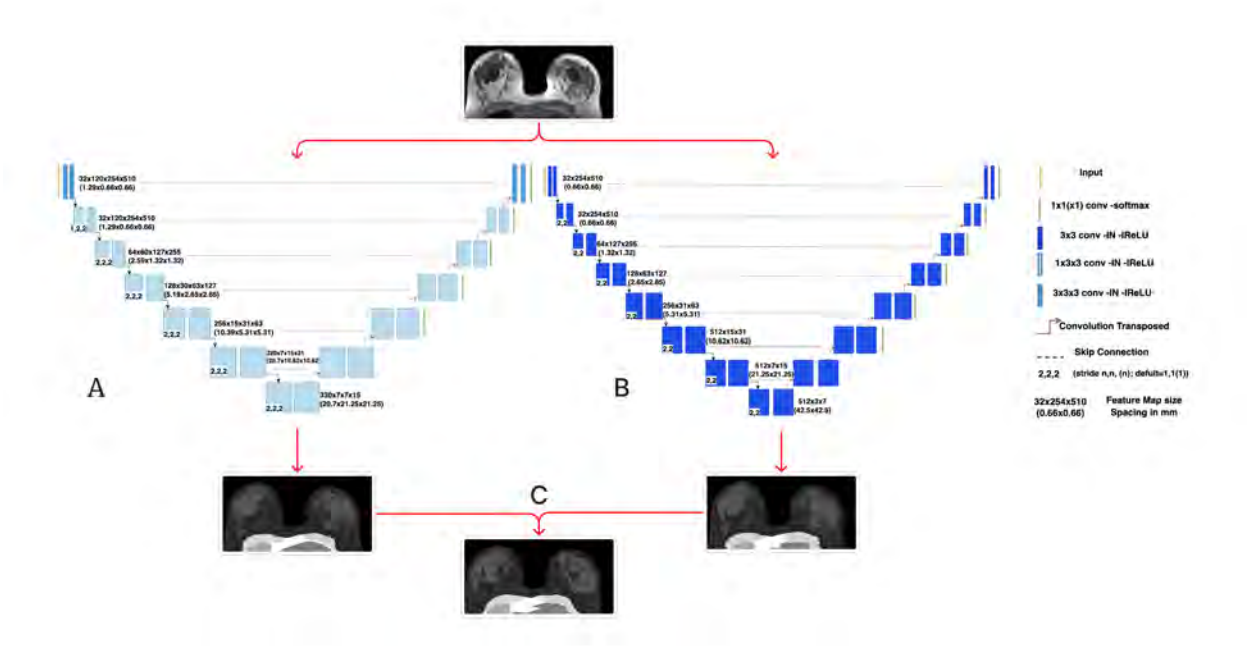



Figure 4: Network architectures generated by nnU-Net for the dataset. 2D and 3D full-resolution and the ensemble of them as a final output.

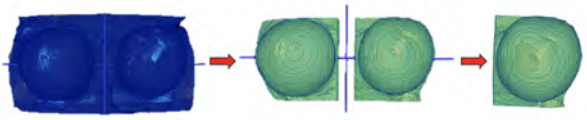


Figure 5: Process of isolating the segmentation map of the breast which will be meshed in the next step.

#### 3.2.3. *Finite Element Analysis: Simulating Compression*

Mammography involves significant breast compression. A Finite element analysis approach is employed to replicate this, dividing the process into 20,000 small steps for stability (Bathe, 2006). The simulation assumes an initial state where the breast is stretched due to patient positioning and gravity during MRI acquisition. The mechanical behavior of the breast under stress is modeled using a Neo-Hookean material model (nearly incompressible, homogeneous, and isotropic) (Wellman, 1999). The model based on NiftySim applies to both glandular and fatty tissues, with specific properties detailed in (Garcia et al., 2017) (Section II-B.3). While skin effects are considered negligible, gravity is included for increased accuracy (Mertzanidou et al., 2014). The model allows for slight breast sliding along the thorax, mimicking the real-life connection via connective tissue. Compression paddles are defined mathematically. From an anatomical perspective, the breast-body connection is not rigid; connective tissue allows some sliding. Therefore, nodes at the breast-body interface can slide parallel to paddle displacement (Mertzanidou et al., 2014). Additionally, the paddles are defined using a planar parametric equation. Since the pectoral muscle and internal organs are excluded, the paddle position relative to the thorax is irrelevant, and the entire breast model is compressed. A frictionless contact model is used between the biomechanical model and paddles. The analysis is performed using NiftySim, a tool designed for soft-tissue simulations (Johnsen et al., 2015). The solver manages various parameters (position, orientation, and elastic properties) to generate both uncompressed and compressed breast models. Initial elastic parameters are based on measurements by (Wellman et al., 1999), with specific values for different tissues (Young's modulus: 4.46 kPa for adipose tissue, 15.1 kPa for glandular tissue; Poisson's ratio: 0.45-0.499 during optimization). Uniform grids are used for spatial indexing to manage elements efficiently. This method transforms physical space coordinates into the internal reference system, ensuring accurate deformation mapping and tracking. High-resolution MRI scans (acquired using a 1.5 Tesla Siemens scanner) provide the foundation for detailed construction and analysis of the biomechanical models.

#### 3.2.4. *Image reconstruction*

Building upon the mesh-based reconstruction technique described in (García et al., 2020), the following steps are undertaken to reconstruct a compressed breast segmentation map voxel data from the biomechanical model. This segmentation map represents the breast tissue in its compressed state, which is crucial for simulating mammographic procedures. A uniform grid is created around the mesh, and each voxel (3D pixel)

within this grid is assigned a specific location based on its relative position to the elements in the mesh. Calculations involving barycentric coordinates are employed to determine a voxel's position within a tetrahedron (element). This essentially entails tracing points from a "ray" in the compressed model back to the original uncompressed MRI data, forming a curve.

After assigning positions to all voxels in the grid, the tissue type (adipose, glandular, or lesion) for each voxel is determined using nearest neighbor interpolation. In simpler terms, the tissue label for a voxel is copied from the closest voxel in the original segmented MRI data. Finally, with all voxels assigned positions and tissue labels, a compressed breast segmentation map is reconstructed. This segmentation map essentially represents the breast tissue after compression, with each voxel containing information about its location and tissue type. The reconstruction process is relatively faster for segmentation maps with larger voxel sizes. However, it can take considerably longer for a segmentation map with very high resolutions (García et al., 2020). The inputs to the stage are the originally generated mesh, the extracted compressed mesh, and the original segmentation map used and the final output is the reconstructed compressed segmentation map.

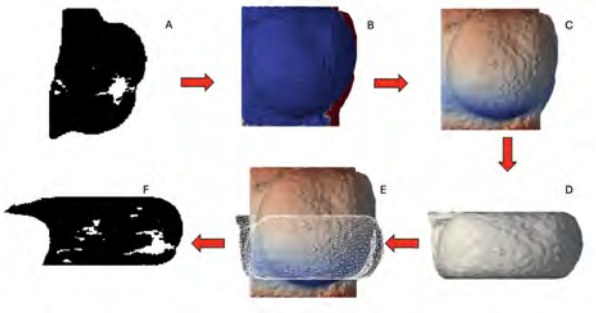



Figure 6: Compressed Breast segmentation map Workflow. (A) Segmentation map of the breast tissue. (B) Generated mesh representing the breast geometry. (C) Displacement magnitudes obtained from NiftySim simulation, visualized as a color map. Red areas represent regions in contact with the compression plates of MRI and experience greater displacement compared to blue areas. (D)Extracted compressed segmentation map (E)For better visualization of compression of breast segmentation map before and after compression (after compression is shown by its wireframe representation) (F) Final reconstructed compressed breast segmentation map.

### *3.3. Lesion inclusion into biomechanical modeling*

This project also explores the inclusion of benign and malignant lesions during the geometry extraction stage. The ultimate goal is to analyze how these lesions affect the biomechanical modeling behavior of the breast. Besides fatty and glandular tissues, other structures can influence the final simulation. These include abnormal tissues with elastic and hyperelastic parameters differing from healthy breast tissue. Notably, the stiffness of these materials varies: for the malignant ones they tend to be stiffer which means that under compression they will be less displaced, while for the benign this value is smaller. The elastic parameters are reported by (Wellman et al., 1999) and (Lorenzen et al., 2002) where in this work the parameter is chosen based on Young's modulus reported in the (Lorenzen et al., 2002) 15.7 kpa for the malignant lesions and the 7 kpa for the benign ones. These values were obtained using magnetic resonance elastography (MRE), a specific form of elastography that leverages magnetic resonance imaging (MRI) to quantify and map the mechanical properties (elasticity or stiffness) of soft tissue (Sarvazyan et al., 1995).

The breast tissue segmentation maps obtained from the nnUnet are integrated with the manually annotated lesion ground truth. This results in the final 3-label segmentation maps, considering the lesion as a third label after fat and glandular tissues. The core objective is to track the biomechanical model's behavior in cases with either benign lesions (water cysts or fibroadenoma) or validated malignant lesions identified by an expert radiologist due to their boundary irregular shapes (examples shown in Figure 2). Figure 7 illustrates the mesh incorporating the lesion.

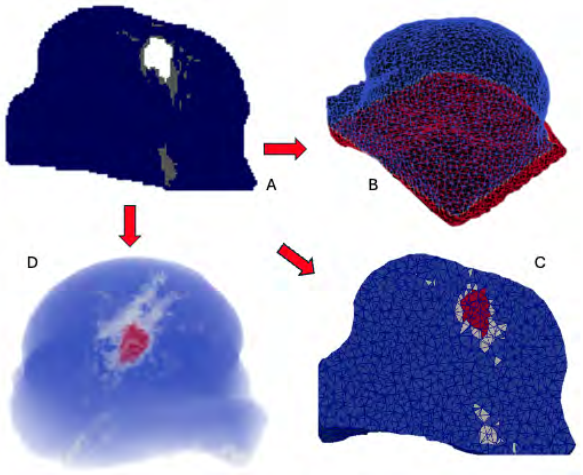



Figure 7: Lesion-Embedded Mesh. (A) Segmentation map incorporating the lesion. (B) Wireframe mesh representation of the segmentation map's boundary condition. (C) Material distribution within the clipped surface edge mesh illustration. (D) Volume rendering of the mesh, highlighting the lesion's location within the glandular tissue.

This study investigates how lesions influence biomechanical models of the breast. We explore three experimental frameworks to achieve this:

- **Baseline Framework (base):**

  Motivation: the initial framework establishes a baseline by analyzing the biomechanical behavior of fatty and glandular tissue excluding lesions. This is crucial for several reasons. First, it allows us to isolate the fundamental biomechanical prop-

erties of these two primary tissue types. Understanding these baseline properties serves as a reference point for subsequent analyses that incorporate lesions. Second, lesions often arise from and share similar characteristics with glandular tissue. By initially combining lesions with glandular tissue, we simplify the model and avoid introducing unnecessary complexity in this foundational stage. However, it is important to acknowledge that some lesions may exhibit distinct mechanical properties. Subsequent frameworks will address this by incorporating lesions as a separate category.

Methodology: the baseline framework utilizes a two-label segmentation map. In this map, areas identified as fatty tissue are retained as a distinct category. Conversely, areas identified as glandular tissue are combined with any identified lesions into a single label. This decision is based on the frequent association of lesions with glandular tissue and their often-similar biomechanical characteristics. By employing this simplified segmentation map, we can isolate the fundamental biomechanical properties of fatty and dense tissue. This initial analysis lays the groundwork for a more nuanced understanding of how lesions may alter these properties in subsequent frameworks.

- **Direct Lesion Inclusion Framework (DLI):**

  Motivation: analyze the model's behavior directly in the presence of the lesion, providing a more realistic representation of the breast tissue.

  Methodology: we employ the segmentation map containing the lesion (three labels) as the initial image for mesh generation. This results in a final reconstructed image and mesh that retains all three distinct labels. Additionally, by assigning specific elastic parameters to each lesion type, we can potentially visualize breast volume reduction across the frameworks.

- **Compressed Lesion Estimation Framework (CLE):**

  Motivation: estimate the lesion's location and its impact on dense tissue volume changes within the context of the two-label framework.

  Methodology: this framework builds upon the two-label framework. However, during compressed image reconstruction, we use the segmentation map containing the lesion (three labels) as input. This preserves all three labels in the compressed image, allowing us to estimate the lesion's location within the two-label framework and observe its influence on dense tissue volume changes.

Figure 8 illustrates the workflow pipelines used for comparing these scenarios.

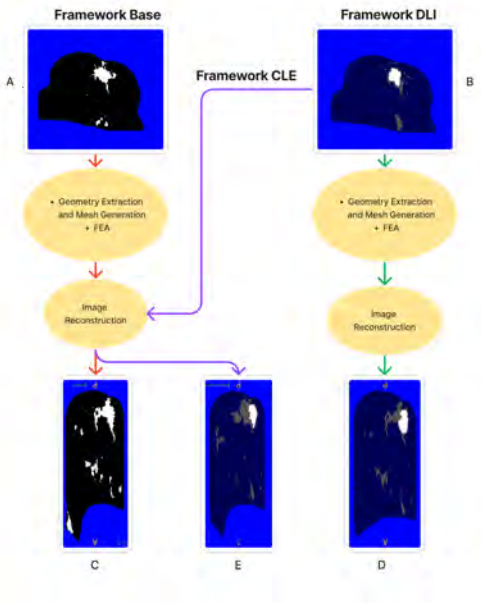



Figure 8: Illustration of 3 experimented frameworks. A) Segmentation map containing fat and glandular tissue, B) Segmentation map containing fat and glandular tissue and lesion, C) Compressed image of Baseline Framework (Base) D) Compressed image of Direct Lesion Inclusion Framework (DLI), E)Compressed image Compressed Lesion Estimation Framework (CLE).

### *3.4. Evaluation*

To assess the effectiveness of our approach, we employed appropriate metrics for both the deep learning segmentation models and the subsequent biomechanical analysis tasks.

- **Segmentation Evaluation:** the Dice coefficient (DSC) (Eelbode et al., 2020), a common metric in image segmentation, was used to evaluate the models' performance. It quantifies the overlap between predicted and ground truth segmentation masks for each tissue class (higher DSC indicates better agreement). Key terms in this context include True Positive (TP), which refers to correctly identified pixels in a specific class, False Positive (FP), which refers to pixels wrongly classified as a specific class, and False Negative (FN), which denotes pixels belonging to a specific class that were missed.

  The formula for DSC is:

$$\mathrm{DSC} = \frac{2\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

- **Biomechanical Evaluation:** We propose using the Dice coefficient to assess lesion shape similarity and evaluate how well the biomechanical

model captures lesion shape changes during compression. The aim is that including lesion material properties in the model will lead to a more accurate representation of the lesion's shape in the compressed breast. This approach builds upon the previously described Dice coefficient, which compares the overlap between lesion regions in the uncompressed and compressed segmentation maps. This metric focuses on the overlap between regions of interest (ROIs) centered around the mass centers of the lesion label in each segmentation. Cubic ROIs were used for consistent analysis, and zero-padding addressed potential spatial dimension differences due to compression. Figure 9 illustrates how lesions are represented and deformed under compression. Additionally, breast tissue loss during compression is assessed by Breast Volume (BV) Change refers to the total tissue volume within a breast, and Volumetric Breast Density (VBD) which quantifies the average density of glandular tissue within the breast. VBD is calculated by dividing the glandular tissue volume (number of voxels labeled 2 multiplied by voxel spacing) by the total breast volume and multiplying by 100%. This method assesses changes in glandular tissue density distribution due to compression, evaluating the model's ability to maintain spatial distribution during simulated compression by comparing VBD values from uncompressed and compressed segmentation maps. The formula for VBD is:

$$\text{VBD} = \left( \frac{\text{Glandular Tissue Volume}}{\text{Breast Volume}} \right) \times 100\%$$

We will present the detailed evaluation results, including qualitative observations and quantitative metrics, in the Experimental Results section.

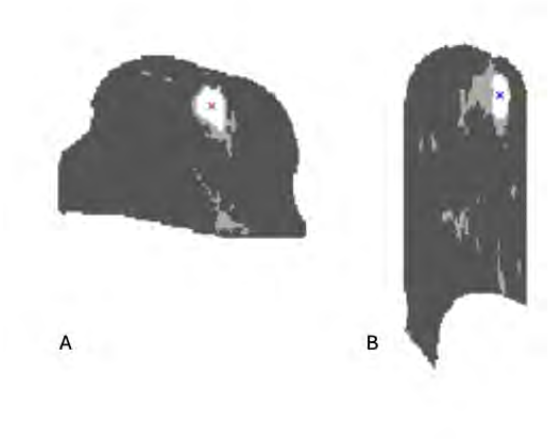



Figure 9: Lesion label mass detected in (A)uncompressed and (B) compressed segmentation map

### *3.5.* ***Implementation details***

A virtual environment based on Python 3.10.11 was used for this project. All other necessary Python libraries were installed from the nnU-Net GitHub repository within this virtual environment, ensuring compatibility and reproducibility. The nnU-Net code utilized in this study is publicly accessible at github.com/MIC-DKFZ/nnUNet.

The deep learning model training and biomechanical modeling simulations were performed using an NVIDIA GeForce RTX 2080 Ti GPU. The environment was configured with PyTorch version 2.1.2+cu121 and CUDA version 12.1. The GPU provided approximately 11 GB of memory, sufficient for handling the complex computations required by the nnU-Net framework. Each fold training for each configuration took around 3 hours, taking up to 14 hours to complete the 5-fold cross-validation with 100 epochs. The GPU capabilities were also utilized to accelerate the compression process using NiftySim, significantly improving computational efficiency.

## 4. Experimental Results

### *4.1.* ***Segmentation Evaluation***

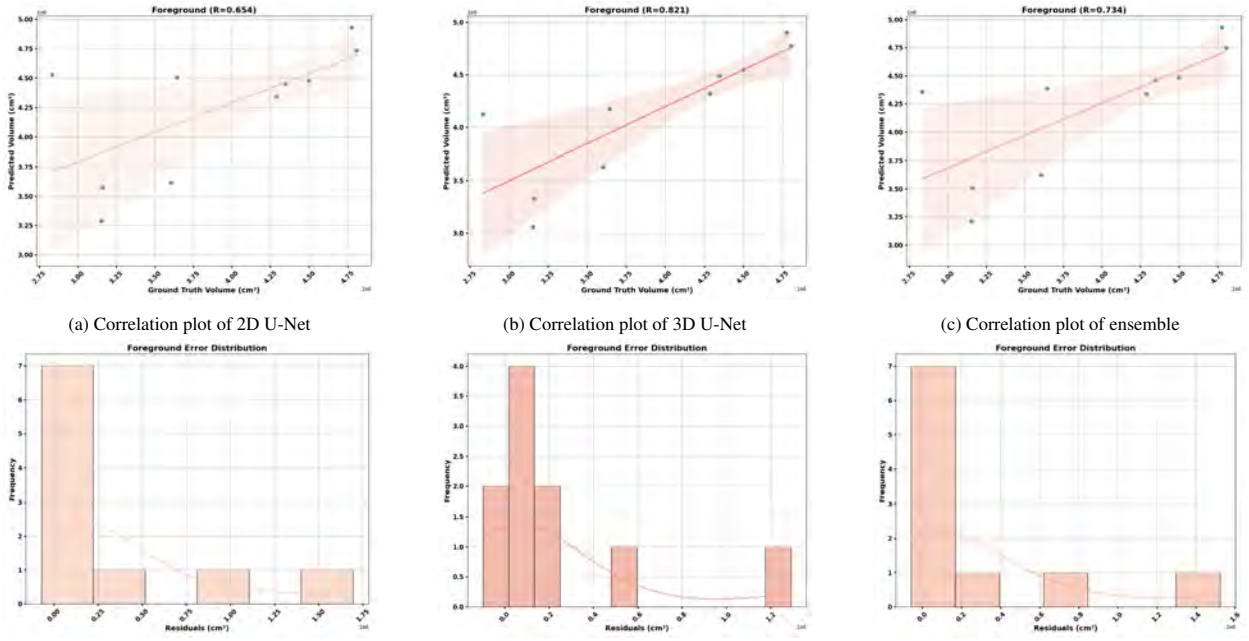
The segmentation evaluation of the inference of the trained model on the 10 specific cases involves a comprehensive comparison of 2D-UNet, 3D-UNet, and their Ensemble methods across various anatomical classes, utilizing both quantitative and qualitative metrics to assess performance.

#### *4.1.1.* ***Quantitative Results***

Table 3 presents the Dice coefficients for different segmentation methods (2D-UNet, 3D-UNet, and their Ensemble) across six classes: Fat, Glandular, Heart, Lung, Pectoral, and Thorax. The Ensemble method consistently shows the highest total mean Dice coefficient of 0.83, indicating its superior performance compared to individual methods. The state-of-the-art methods ((Alqaoud et al., 2022a); (Zafari et al., 2019); (Alqaoud et al., 2022b)), however, surpass these methods in individual classes but lack some comprehensive data which includes the segmentation of the organs included in the breast MRI for a complete comparison. Key observations include the 2D-UNet performing best in the Fat class with a Dice score of 0.94, the 3D-UNet showing slightly better performance in the Heart class with a Dice score of 0.79, and the Ensemble method achieving the highest total mean Dice coefficient, suggesting that combining 2D and 3D approaches leverages the strengths of both. Notably, this work, to the best of our knowledge, is the first to perform segmentation on real breast MRI including all organs, not just the breast region, highlighting its significance and positive aspect. Building on the promising results from the Dice

| Methods | Fat | Glandular | Heart | Lung | Pectoral | Thorax | Total Mean |
|---|---|---|---|---|---|---|---|
| 2D-UNet | 0.94 | 0.88 | 0.77 | 0.72 | 0.87 | 0.72 | 0.82 |
| 3D-UNet | 0.93 | 0.86 | 0.79 | 0.72 | 0.85 | 0.72 | 0.81 |
| Ensemble | **0.94** | **0.88** | **0.79** | **0.73** | **0.87** | **0.74** | **0.83** |
| State of the Art | 0.95 | 0.87 | - | - | 0.89 | - | - |

Table 3: Comparison of Different Methods across various Segmented Classes with state of the art((Alqaoud et al., 2022a); (Zafari et al., 2019); (Alqaoud et al., 2022b))




(a) Correlation plot of 2D U-Net

(b) Correlation plot of 3D U-Net

(c) Correlation plot of ensemble

(d) Error distribution plot of 2D U-Net

(e) Error distribution plot of 3D U-Net

(f) Error distribution plot of ensemble

Figure 10: Correlation (top row) and error distribution (bottom row) for 2D U-Net, 3D U-Net, and ensemble (R-values: 'coolwarm' colormap, red=stronger correlation). Dice (0.822, 0.819, 0.83) and R (0.654, 0.820, 0.734) values are shown. The ensemble exhibits improved spatial overlap and volume consistency.

coefficient analysis, we conducted a correlation analysis to evaluate the relationship between the predicted and ground truth volumes for each segmentation method Figure 10. While the 3D U-Net achieved a slightly higher correlation coefficient (R = 0.820) for individual classes (Figure 8b), the ensemble method exhibited a strong correlation (R = 0.734) (Figure 8c). This suggests that the ensemble method prioritizes capturing the overall relationship between the predicted probabilities and the actual volumes across all classes. This focus on the broader trend translates to more consistent predictions, as evidenced by the narrower error distribution in the ensemble method (Figure 8f) compared to 2D U-Net (Figure 8d) and 3D U-Net (Figure 8e). These results, along with the high Dice coefficients, indicate that the ensemble method not only predicts accurate segmentation masks but also captures the underlying relationships between the data and the labels in a way that generalizes better across all organ classes.

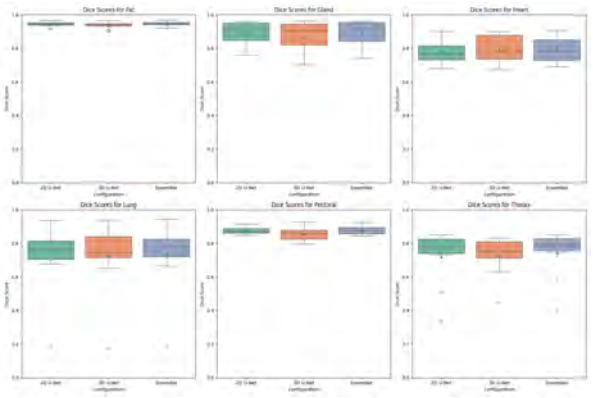


Figure 11: Box plot of 3 models indicating the mean and variance of each class in the 2 configurations and their ensemble

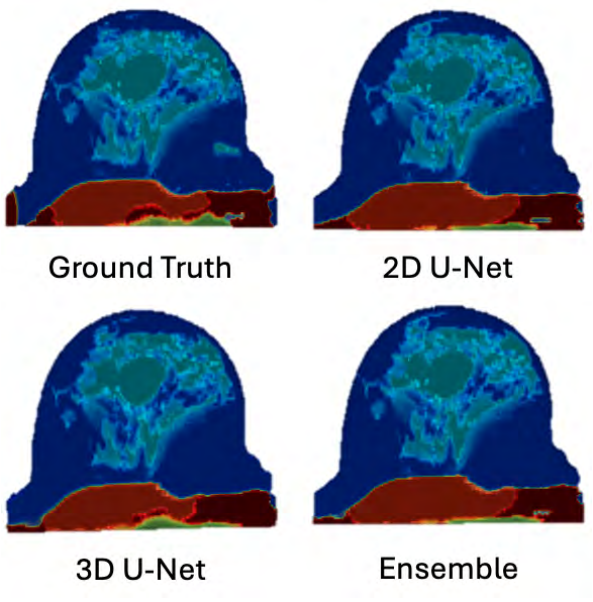



Figure 12: Clipped 3D view of segmentation result in JET color map. The color map assigns dark blue to fat, light blue to glandular tissue, green to the heart, orange to the pectoral muscle, and brown to the thorax. In slice 60, the lung is not shown.

Figure 11 presents the distribution of Dice coefficients across different segmentation methods for each class. We observe that the Ensemble method generally achieves the highest median Dice coefficients across most classes, particularly for Fat. This aligns with the findings in Table 3, where the ensemble method showed the highest overall mean Dice coefficient. The boxplots also reveal that the interquartile range (IQR) and whisker lengths are generally smaller for the ensemble method compared to other methods for several classes (e.g., pectoral), indicating more consistent performance. However, some outliers are present in the Lung and Thorax class for all methods, suggesting potential challenges in segmenting this particular class.

#### *4.1.2. **Qualitative Results***

Figure 13 offers a detailed qualitative assessment of segmentation results for a patient from a test set with a large breast shape and moderately dense glandular tissue from the test set. The figure includes axial, sagittal, and coronal MRI slices to provide a comprehensive three-dimensional view of the segmented structures. The top row displays the original MRI images, while the bottom row depicts the ground truth segmentation (ideal segmentation), predictions by 2D U-Net, 3D U-Net, and the ensemble method, respectively. The color scheme assigns black to the background, dark blue to the fat, light blue to glandular tissue, green to the heart, yellow to the lung, orange to the pectoral muscle, and brown to the thorax.

A close examination reveals generally good agreement between the predicted segmentation masks and the ground truth for most tissues. However, a closer inspection might identify potential discrepancies in the segmentation of specific organs or tissues across different slices. For example, in the coronal slice, some models might struggle to differentiate between dense glandular tissue and surrounding structures. We can further evaluate the performance of distinct tissue types. For instance, some tissues like fat might be consistently well-segmented across all methods, indicating robust performance. Conversely, other tissues like dense glandular tissue might pose challenges, particularly for differentiating them from neighboring structures.

The comparison can also reveal potential advantages of the ensemble method. By analyzing regions with subtle tissue boundaries, we can determine if the ensemble method offers a more accurate delineation compared to 2D U-Net and 3D U-Net. This analysis can highlight the ensemble's ability to capture complex tissue relationships, potentially leading to improved segmentation accuracy. This qualitative assessment, alongside the quantitative metrics presented earlier (e.g., Dice coefficient), provides a richer understanding of each model's strengths and limitations. This combined analysis offers valuable insights into the effectiveness of different segmentation approaches for real-world medical image analysis applications.

For better visualization, the Figure 12 clipped 3D view of the segmentation result of another patient from the test set has been provided which shows more details and the strengths and weaknesses of each network's capabilities.

### *4.2. **Biomechanical Evaluation***

Facing the limitations of the NiftySim from 10 cases we were able to perform compression on four of them. Hence, the focus in this part will be to assess the two frameworks for four cases whose outputs contain lesion labels (DLI and CLE) in their compressed segmentation maps to achieve our goal, even though we had the output for the baseline which contains 2 labels. Framework DLI and Framework CLE are evaluated through a combination of quantitative metrics and qualitative visual assessments. This comprehensive comparison illustrates the effectiveness of each framework in accurately modeling the biomechanical behavior of breast tissue with included lesions. We also propose to use Dice coefficients 3.4 to assess the shape changes of the lesion region with and without taking into account the lesion material properties in the biomechanical model. The aim is that by including lesion information, the shape of the lesion in the compressed breast will be more similar to the original uncompressed volume.

#### *4.2.1. **Quantitative Results***

Leveraging the segmentation results obtained from the nnU-Net and overlaying their lesion ground truth for

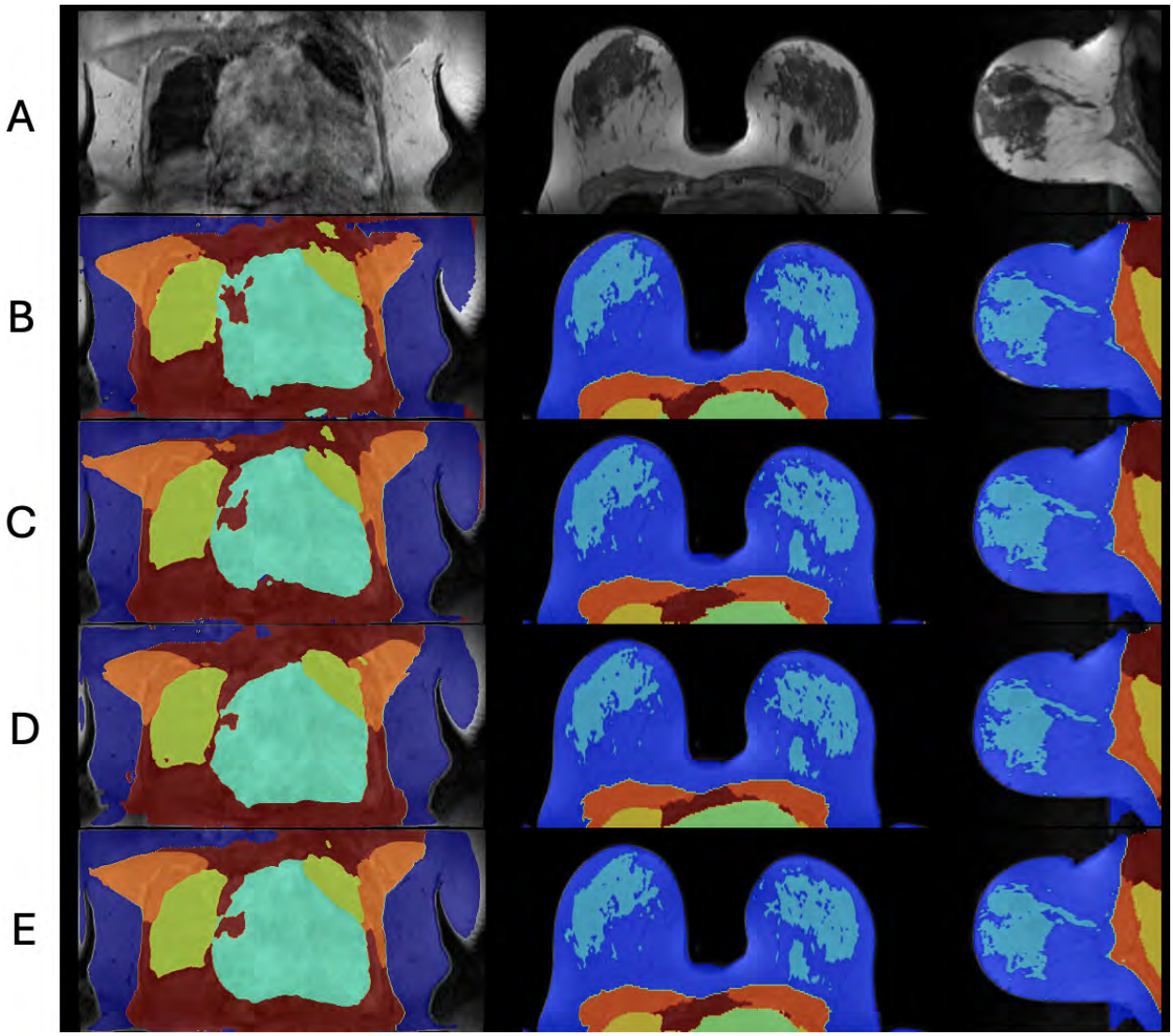



Figure 13: Segmentation result of the (A) MRI image, (B) Ground Truth, (C) 2D U-Net prediction, (D) 3D-UNet prediction, (E) Ensemble of 2D and 3D U-Net

further biomechanical modeling, we assessed the comparison of the methods for the presence of lesions in the framework. The quantitative evaluation compares the accuracy and effectiveness of Framework DLI and Framework CLE in modeling the biomechanical behavior of breast tissue with included lesions. Various metrics such as Dice coefficients, breast volume changes, and volumetric breast density are analyzed to assess the performance of each framework.

Figure 14 illustrates the Dice coefficients calculated to quantify the overlap between the compressed and uncompressed lesion segmentation map. Specifically, Dice-CLE measures the similarity between the uncompressed and CLE compressed segmentation map, Dice-DLI evaluates the overlap between uncompressed and DLI compressed segmentation maps, and Dice-DLI-CLE assesses the overlap for each tissue class (Fat, Glandular tissue, and Lesion) between the compressed segmentation map of the two frameworks DLI and CLE to assess how much the output of DLI and CLE is overlapped. Higher Dice values indicate better high overlap. By using Dice coefficients to assess the shape changes of the lesion region, we can evaluate how well the biomechanical model preserves the lesion's shape under compression. We hypothesize that including lesion information will result in a compressed lesion shape that is more similar to the original uncompressed volume, thereby validating the effectiveness of the model.

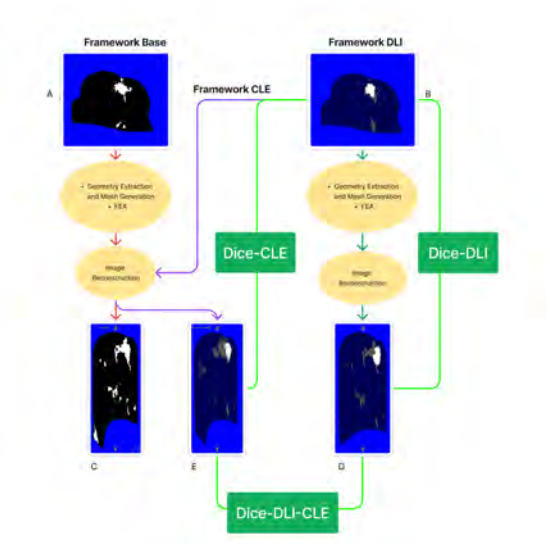



Figure 14: The dices that are calculated to assess the effect of the lesion label in each framework

Table 4a presents the dice coefficients (Dice-DLI and Dice-CLE). This suggests that Framework DLI is more effective in maintaining the accuracy of lesion segmentation under compression which can be expected since the lesion material properties are included initially so it is more preserved. Table 4b The tables also summarize the percentage changes in breast volume (BV) and volumetric breast density (BVD) for each framework across different cases. These metrics provide insights into the effects of compression on various tissue types in both DLI (Direct Lesion Inclusion) and CLE (Compressed Lesion Estimation) frameworks. For fat tissue, the BV reduction in both DLI and CLE is consistent for each case, showcasing that both frameworks handle fat tissue compression similarly. Regarding glandular tissue for each case, the BV reduction in DLI demonstrates no direct correlation between glandular tissue loss and lesion loss. This is because different elastic parameters are assigned to these tissues in the biomechanical model, causing them to behave differently under compression. Conversely, in CLE, there is a noticeable loss in both glandular tissue and lesions. This indicates that the framework does not distinctly preserve the properties of each tissue type, leading to more significant tissue loss overall. For lesion tissues for each case, the BV reduction in DLI is slightly lower because lesion-specific elastic parameters are assigned from the beginning. This ensures that during compression, the biomechanical models consider the unique properties of lesions. The data shows that malignant cases experience less tissue loss than benign cases in this framework due to the initial integration of lesion-specific parameters. In CLE, however, the BV reduction for lesions is more significant compared to glandular tissue. This suggests that the properties of lesions are not preserved as effectively, leading to greater tissue loss. Notably, this loss is more pronounced in benign cases than malignant ones, as malignant tissues are more resistant to displacement. Generally, the BV changes measure tissue loss during compression, while the BVD assesses changes in the distribution of glandular tissue. The DLI framework shows slightly more consistent BV and BVD values, indicating it provides a more stable and accurate biomechanical model when lesions are included from the beginning. This consistency suggests that DLI is better at preserving tissue properties and accurately modeling breast tissue behavior under compression.

Furthermore, Table 4c provides the Dice-DLI-CLE coefficients for each tissue class (Fat, Glandular tissue, and Lesion) between the compressed segmentation maps of Framework DLI and Framework CLE. Higher Dice-DLI-CLE values indicate how each tissue overlaps in both compressed segmentation maps of DLI and CLE and the ability to maintain accurate tissue class segmentation under compression, emphasizing its effectiveness in modeling both benign and malignant lesions.

#### *4.2.2. **Qualitative Results***

Figure 8 presents the qualitative results of biomechanical modeling for the three frameworks of base, DLI, and CLE. It includes segmentation maps and compressed segmentation maps for each framework. The comparison shows that Framework DLI captures lesion boundaries and properties more distinctly and accurately integrates lesions within the biomechanical model, resulting in more realistic tissue deformation behavior.

## 5. Discussion

This study investigates the impact of lesion inclusion on biomechanical modeling of breast tissue, utilizing advanced segmentation techniques provided by the nnU-Net framework. The key question addressed is whether including lesions from the start (Framework DLI) results in more accurate and realistic biomechanical models compared to estimating lesion impacts post-compression (Framework CLE). This research builds on existing literature by incorporating lesion-specific mechanical properties into biomechanical models, an area previously underexplored.

Our findings demonstrate that Framework DLI, which includes lesions from the outset, offers several advantages. Quantitative analysis reveals higher Dice coefficients for Framework DLI, indicating better overlap and representation of lesions. This result aligns with previous studies that emphasize the importance of accurate initial conditions in modeling complex biological systems. Framework DLI also shows more consistent and realistic changes in breast volume and volumetric breast density, which are crucial for simulating the mechanical behavior of breast tissue under compression.

Table 4: Comparison of methods for different cases

| Methods | Case 1 (Benign) | Case 2 (Benign) | Case 3 (malignant) | Case 4 (malignant) |
|---|---|---|---|---|
| Dice-DLI | 0.6909 | 0.7116 | 0.6376 | 0.6645 |
| Dice-CLE | 0.6824 | 0.7133 | 0.6407 | 0.6656 |

(a) Dice-DLI and Dice-CLE which for lesion label between the compressed segmentation map and uncompressed segmentation map of a framework with direct lesion inclusion and estimation of lesion

| Methods | Case 1 | Case 2 | Case 3 | Case 4 |
|---|---|---|---|---|
| **BV-DLI** | Fat:1.56%<br>Gland:0.42%<br>Lesion:4.60% | Fat:1.95%<br>Gland:0.28%<br>Lesion:5.26% | Fat:1.97%<br>Gland:0.07%<br>Lesion:2.63% | Fat:1.86%<br>Gland:0.07%<br>Lesion:2.94% |
| **Total BV** | 1.56% | 1.88% | 1.94% | 1.61% |
| **BVD** | 1.72%-1.72% | 3.54%-3.61% | 1.97%-2.01% | 13.87%-14.09% |
| **BV-CLE** | Fat: 1.55%<br>Gland: 0.28%<br>Lesion:10.34% | Fat:1.94%<br>Gland:0.07%<br>Lesion:5.74% | Fat:1.98%<br>Gland:0.08%<br>Lesion:0.88% | Fat:1.86%<br>Gland:0.05%<br>Lesion:1.63% |
| **Total BV** | 1.53% | 1.87% | 1.94% | 1.61% |
| **BVD** | 1.72%-1.75% | 3.54%-3.61% | 1.97%-2.01% | 13.87%-14.09% |

(b) BV and BVD for each framework DLI and CLE for malignant and benign

| Classes | Fat | Glandular tissue | Lesion |
|---|---|---|---|
| **Case 1** | 0.7578 | 0.9925 | 0.9997 |
| **Case 2** | 0.9841 | 0.9879 | 0.9994 |
| **Case 3** | 0.9828 | 0.9907 | 0.9997 |
| **Case 4** | 0.9904 | 0.9959 | 0.9993 |

(c) Dice-DLI-CLE for each label between the compressed segmentation map of framework DLI and CLE

The qualitative assessments further support the superiority of Framework DLI. Visual inspections indicate that Framework DLI provides a seamless integration of lesions within the biomechanical model, resulting in smoother and more realistic deformation behavior. This is particularly significant for malignant lesions, which are stiffer and less displaceable. Including these lesions from the start ensures their mechanical properties are accurately modeled throughout the compression process, leading to better predictions of their behavior and impact on surrounding tissues. This finding is consistent with the literature that underscores the need for precise mechanical property assignment in biomechanical simulations.

In contrast, Framework CLE, which estimates lesion impacts post-compression, is less effective in capturing the true mechanical behavior of lesions. Although useful for quick estimations, Framework CLE may introduce artifacts or inconsistencies in the reconstructed compressed images. This limitation is especially evident in scenarios involving small breast sizes or poor image quality, such as those encountered in the TCGA-BRCA dataset ((Burnside et al., 2016); (Clark et al., 2013). Our attempts to leverage this dataset were hampered by the low quality and limited availability of mammograms, further highlighting the need for high-quality datasets in biomechanical modeling.

The implementation of deep learning techniques, specifically the nnU-Net framework, proved beneficial for automating tissue segmentation. Notably, this study is among the first to perform segmentation on real-world breast MRI data that includes other organs, making the research particularly challenging. This approach not only enhanced segmentation accuracy but also facilitated the incorporation of lesion data into the biomechanical models. The nnU-Net's ability to automatically configure its architecture based on the dataset reduced the need for manual adjustments and ensured efficient processing of diverse breast DCE-MRI datasets. This innovation is significant as it bridges the gap between advanced medical imaging and practical clinical applications, potentially leading to improved outcomes for breast cancer patients.

However, the study also faced significant challenges. Among the 10 cases that were supposed to be analyzed for lesion inclusion, the NiftySim framework did not operate on 4 of them, and in 2 cases, the compression was incorrect, producing abnormal results. Furthermore, the finite element analysis using Febio failed for all cases due to mesh quality issues. These issues highlight the sensitivity of the NiftySim framework to factors like the distance of the nipple to the chest wall and underscore the limitations posed by mesh quality in Febio. These challenges point to an essential area for future research and development.

## 6. Future directions

Future research should focus on addressing the identified challenges by improving the robustness and accuracy of biomechanical modeling frameworks. Enhancing the mesh quality and resolving the issues associated with the NiftySim framework could lead to more reliable simulations. Additionally, developing advanced preprocessing techniques to handle low-quality or incomplete datasets, like those from the TCGA-BRCA, would make the modeling process more resilient. Integrating more sophisticated deep learning algorithms with biomechanical models could further refine lesion segmentation and mechanical property assignment. Specifically, applying deep learning techniques to finite element analysis (FEA) could revolutionize the accuracy and efficiency of these simulations. Expanding the number of cases analyzed in future studies would also enhance the generalizability and robustness of the findings. Collaborating with clinical partners to access higher quality and larger datasets would be invaluable. Moreover, exploring the application of these improved models in clinical settings to assess their potential to enhance breast cancer diagnosis and treatment planning should be a priority. These efforts will not only advance the field of biomechanical modeling but also improve patient outcomes by providing more accurate and reliable tools for clinical use.

## 7. Conclusions

This study demonstrates the significant advantages of including lesions from the outset in the biomechanical modeling of breast tissue. Utilizing the nnU-Net framework for segmentation, we achieved higher accuracy in tissue delineation and successfully incorporated lesion-specific mechanical properties into the models. Framework DLI (Direct Lesion Inclusion) consistently outperformed Framework CLE (Compressed Lesion Estimation) in both quantitative and qualitative assessments, providing more accurate and realistic simulations of tissue deformation. Our findings underscore the importance of accurate initial conditions and robust segmentation techniques in biomechanical simulations. Despite facing challenges with dataset quality and framework sensitivity, this research advances the integration of deep learning with biomechanical modeling, offering valuable insights for improving clinical outcomes in breast cancer treatment.

## Acknowledgments

This project would not have reached its full potential without the guiding light of Professor Robert Marti, whose supervision steered me in the right direction. My co-supervisor, Eloy Garcia Marcos, provided invaluable support throughout the journey. A debt of gratitude is owed to Professor Kai Villanova for the diligent validation of my manual annotations. Special thanks to my friend Hadeel Awwad, whose unwavering support and willingness to integrate her FEBio model with my data were instrumental in enriching the evaluation process. Finally, to my ever-supportive family, your love and encouragement fueled my perseverance.

## References

Alqaoud, M., Plemmons, J., Feliberti, E., Kaipa, K., Dong, S., Fichtinger, G., Xiao, Y., Audette, M., 2022a. Multi-modality breast mri segmentation using nnu-net for preoperative planning of robotic surgery navigation, in: 2022 Annual Modeling and Simulation Conference (ANNSIM), IEEE. pp. 317–328.

Alqaoud, M., Plemmons, J., Feliberti, E., Kaipa, K., Dong, S., Fichtinger, G., Xiao, Y., Audette, M., 2022b. Multi-modality breast mri segmentation using nnu-net for preoperative planning of robotic surgery navigation, in: 2022 Annual Modeling and Simulation Conference (ANNSIM), pp. 317–328. doi:10.23919/ANNSIM55834.2022.9859361.

Arlinghaus, L.R., Welch, E.B., Chakravarthy, A.B., Xu, L., Farley, J.S., Abramson, V.G., Grau, A.M., Kelley, M.C., Mayer, I.A., Means-Powell, J.A., et al., 2011. Motion correction in diffusion-weighted mri of the breast at 3t. Journal of Magnetic Resonance Imaging 33, 1063–1070.

Babarenda Gamage, T.P., Rajagopal, V., Nielsen, P.M., Nash, M.P., 2012. Patient-specific modeling of breast biomechanics with applications to breast cancer detection and treatment. Patient-Specific Modeling in Tomorrow's Medicine , 379–412.

Bathe, K.J., 2006. Finite element procedures. Klaus-Jurgen Bathe.

Burnside, E.S., Drukker, K., Li, H., Bonaccio, E., Zuley, M., Ganott, M., Net, J.M., Sutton, E.J., Brandt, K.R., Whitman, G.J., et al., 2016. Using computer-extracted image phenotypes from tumors on breast magnetic resonance imaging to predict breast cancer pathologic stage. Cancer 122, 748–757.

Chung, J.H., Rajagopal, V., Nielsen, P.M., Nash, M.P., 2008. Modelling mammographic compression of the breast, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008: 11th International Conference, New York, NY, USA, September 6-10, 2008, Proceedings, Part II 11, Springer. pp. 758–765.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al., 2013. The cancer imaging archive (tcia): maintaining and operating a public information repository. Journal of digital imaging 26, 1045–1057.

Dalmış, M.U., Litjens, G., Holland, K., Setio, A., Mann, R., Karssemeijer, N., Gubern-Mérida, A., 2017. Using deep learning to segment breast and fibroglandular tissue in mri volumes. Medical physics 44, 533–546.

Del Palomar, A.P., Calvo, B., Herrero, J., López, J., Doblaré, M., 2008. A finite element model to accurately predict real deformations of the breast. Medical engineering & physics 30, 1089–1097.

Eelbode, T., Bertels, J., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R., Blaschko, M.B., 2020. Optimization for medical image segmentation: theory and practice when evaluating with dice score or jaccard index. IEEE transactions on medical imaging 39, 3679–3690.

van Engeland, S., Snoeren, P., Hendriks, J., Karssemeijer, N., 2003. A comparison of methods for mammogram registration. IEEE Transactions on Medical Imaging 22, 1436–1444.

Garcia, E., Diez, Y., Diaz, O., Llado, X., Gubern-Mérida, A., Marti, R., Marti, J., Oliver, A., 2017. Multimodal breast parenchymal patterns correlation using a patient-specific biomechanical model. IEEE transactions on medical imaging 37, 712–723.

García, E., Diez, Y., Diaz, O., Lladó, X., Martí, R., Martí, J., Oliver, A., 2018. A step-by-step review on patient-specific biomechanical finite element models for breast mri to x-ray mammography registration. Medical physics 45, e6–e31.

García, E., Fedon, C., Caballo, M., Martí, R., Sechopoulos, I., Diaz, O., 2020. Realistic compressed breast phantoms for medical physics applications, in: 15th International Workshop on Breast Imaging (IWBI2020), SPIE. pp. 30–37.

García, E., Oliver, A., Díaz, O., Diez, Y., Gubern-Mérida, A., Martí, R., Martí, J., 2017. Mapping 3d breast lesions from full-field digital mammograms using subject-specific finite element models, in: Medical Imaging 2017: Image-Guided Procedures, Robotic Interventions, and Modeling, SPIE. pp. 26–33.

García, E., Diez, Y., Diaz, O., Lladó, X., Gubern-Mérida, A., Martí, R., Martí, J., Oliver, A., 2018. Multimodal breast parenchymal patterns correlation using a patient-specific biomechanical model. IEEE Transactions on Medical Imaging 37, 712–723. doi:10.1109/TMI.2017.2749685.

Giess, C.S., Yeh, E.D., Raza, S., Birdwell, R.L., 2014. Background parenchymal enhancement at breast mr imaging: normal patterns, diagnostic challenges, and potential for false-positive and false-negative interpretation. Radiographics 34, 234–247.

Gubern-Mérida, A., Kallenberg, M., Martí, R., Karssemeijer, N., 2012. Segmentation of the pectoral muscle in breast mri using atlas-based approaches, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012: 15th International Conference, Nice, France, October 1-5, 2012, Proceedings, Part II 15, Springer. pp. 371–378.

Gubern-Mérida, A., Kallenberg, M., Mann, R.M., Martí, R., Karssemeijer, N., 2015. Breast segmentation and density estimation in breast mri: A fully automatic framework. IEEE Journal of Biomedical and Health Informatics 19, 349–357. doi:10.1109/JBHI.2014.2311163.

Gubern-Mérida, A., Kallenberg, M., Martí, R., Karssemeijer, N., 2012. Segmentation of the pectoral muscle in breast mri using atlas-based approaches, pp. 371–8. doi:10.1007/978-3-642-33418-4_46.

Huo, L., Hu, X., Xiao, Q., Gu, Y., Chu, X., Jiang, L., 2021. Segmentation of whole breast and fibroglandular tissue using nnu-net in dynamic contrast enhanced mr images. Magnetic Resonance Imaging 82, 31–41.

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.

Johnsen, S.F., Taylor, Z.A., Clarkson, M.J., Hipwell, J., Modat, M., Eiben, B., Han, L., Hu, Y., Mertzanidou, T., Hawkes, D.J., et al., 2015. Niftysim: A gpu-based nonlinear finite element package for simulation of soft tissue biomechanics. International journal of computer assisted radiology and surgery 10, 1077–1095.

Lorenzen, J., Sinkus, R., Lorenzen, M., Dargatz, M., Leussler, C., Röschmann, P., Adam, G., 2002. Mr elastography of the breast: preliminary clinical results, in: RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren, © Georg Thieme Verlag Stuttgart· New York. pp. 830–834.

Melbourne, A., Cahill, N.D., Tanner, C., Hawkes, D.J., 2011. Image registration using an extendable quadratic regulariser, in: 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, IEEE. pp. 557–560.

Mertzanidou, T., Hipwell, J., Johnsen, S., Han, L., Eiben, B., Taylor, Z., Ourselin, S., Huisman, H., Mann, R., Bick, U., et al., 2014. Mri to x-ray mammography intensity-based registration with simultaneous optimisation of pose and biomechanical transformation parameters. Medical image analysis 18, 674–683.

Mertzanidou, T., et al., 2011. Finite element modeling of breast tissue deformation for imaging applications. Medical Image Analysis .

Müller-Franzes, G., Müller-Franzes, F., Huck, L., Raaff, V., Kemmer, E., Khader, F., Arasteh, S.T., Lemainque, T., Kather, J.N., Nebelung, S., et al., 2023. Fibroglandular tissue segmentation in breast mri using vision transformers: a multi-institutional evaluation. Scientific Reports 13, 14207.

del Palomar, A., et al., 2008. Finite element simulation and validation of breast deformation during compression. Medical Physics .

Pathmanathan, P., et al., 2008. Predicting tumor location by modeling the deformation of the breast. IEEE Transactions on Biomedical Engineering .

Pinto Pereira, S.M., Hipwell, J.H., McCormack, V.A., Tanner, C., Moss, S.M., Wilkinson, L.S., Khoo, L.A., Pagliari, C., Skippage, P.L., Kliger, C.J., et al., 2010. Automated registration of diagnostic to prediagnostic x-ray mammograms: Evaluation and comparison to radiologists' accuracy. Medical physics 37, 4530–4539.

Razavi, M., Wang, L., Gubern-Mérida, A., Ivanovska, T., Laue, H., Karssemeijer, N., Hahn, H.K., 2015. Towards accurate segmentation of fibroglandular tissue in breast mri using fuzzy c-means and skin-folds removal, in: Image Analysis and Processing—ICIAP 2015: 18th International Conference, Genoa, Italy, September 7-11, 2015, Proceedings, Part I 18, Springer. pp. 528–536.

Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., Hawkes, D.J., 1999. Nonrigid registration using free-form deformations: application to breast mr images. IEEE transactions on medical imaging 18, 712–721.

Sarvazyan, A., Skovoroda, A., Emelianov, S., Fowlkes, J., Pipe, J., Adler, R., Buxton, R., Carson, P., 1995. Biophysical bases of elasticity imaging. Acoustical imaging , 223–240.

Schlömer, N., 2021. pygalmesh: Python interface for CGAL's meshing tools. URL: https://doi.org/10.5281/zenodo.5628848, doi:10.5281/zenodo.5628848.

Siegler, P., Ebrahimi, M., Holloway, C.M., Thevathasan, G., Plewes, D.B., Martel, A., 2012. Supine breast mri and assessment of future clinical applications. European journal of radiology 81, S153–S155.

Smith, T.J., 2013. Breast cancer surveillance guidelines. Journal of oncology practice 9, 65.

Solves-Llorens, J., et al., 2014a. A complete software application for automatic registration of x-ray mammography and magnetic resonance images. Medical Physics .

Solves-Llorens, J.A., Rupérez, M., Monserrat, C., Feliu, E., García, M., Lloret, M., 2014b. A complete software application for automatic registration of x-ray mammography and magnetic resonance images. Medical physics 41, 081903.

The CGAL Project, 2024. CGAL User and Reference Manual. 5.6.1 ed., CGAL Editorial Board. URL: https://doc.cgal.org/5.6.1/Manual/packages.html.

Vidal, J., Vilanova, J.C., Martí, R., et al., 2022. A u-net ensemble for breast lesion segmentation in dce mri. Computers in Biology and Medicine 140, 105093.

Wellman, P., 1999. Tactile Imaging. Harvard University Ph. D. Ph.D. thesis. thesis.

Wellman, P., Howe, R.D., Dalton, E., Kern, K.A., 1999. Breast tissue stiffness in compression is correlated to histological diagnosis. Harvard BioRobotics Laboratory Technical Report 1.

Wu, S., Weinstein, S., Keller, B.M., Conant, E.F., Kontos, D., 2012a. Fully-automated fibroglandular tissue segmentation in breast mri, in: Breast Imaging: 11th International Workshop, IWDM 2012, Philadelphia, PA, USA, July 8-11, 2012. Proceedings 11, Springer. pp. 244–251.

Wu, S., Weinstein, S., Kontos, D., 2012b. Atlas-based probabilistic fibroglandular tissue segmentation in breast mri, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012: 15th International Conference, Nice, France, October 1-5, 2012, Proceedings, Part II 15, Springer. pp. 437–445.

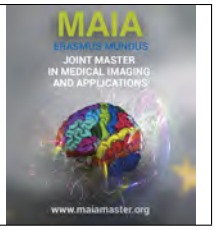
Zafari, S., Diab, M., Eerola, T., Hanson, S.E., Reece, G.P., Whitman, G.J., Markey, M.K., Ravi-Chandar, K., Bovik, A., Kälviäinen, H., 2019. Automated segmentation of the pectoral muscle in axial breast mr images, in: Advances in Visual Computing: 14th International Symposium on Visual Computing, ISVC 2019, Lake Tahoe, NV, USA, October 7–9, 2019, Proceedings, Part I 14, Springer. pp. 345–356.

Zhang, Y., Chen, J.H., Chang, K.T., Park, V.Y., Kim, M.J., Chan, S., Chang, P., Chow, D., Luk, A., Kwong, T., et al., 2019. Automatic breast and fibroglandular tissue segmentation in breast mri using deep learning by a fully-convolutional residual neural network u-net. Academic radiology 26, 1526–1535.

Zhang, Y., Qiu, Y., Goldgof, D.B., Sarkar, S., Li, L., 2007. 3d finite element modeling of nonrigid breast deformation for feature registration in-ray and mr images, in: 2007 IEEE Workshop on Applications of Computer Vision (WACV'07), IEEE. pp. 38–38.

Medical Imaging and Applications

Master Thesis, June 2024




# Deep Learning-Driven Automated Segmentation in High-Resolution 3D Histological Mouse Brain Imaging

Taiabur Rahman[a,b], Alain Lalande[a], Binnaz Yalcin[b], Fabrice Meriaudeau[a], Stephan Collins[b]

[a]*ICMUB Laboratory, CNRS UMR 6302, University of Burgundy, 21078 Dijon, France*
[b]*NeuroGeMMLaboratory, INSERM Unit 1231, University of Burgundy, 21078 Dijon, France*

**Abstract**

The study of the mouse brain is crucial in neuroscience as it serves as an optimal model for understanding the human brain. Genetic manipulation in mice enables exploration of gene effects on brain development. A recent study from the host laboratory identified 198 genes through high-throughput preclinical studies using detailed histological images of thousands of mouse brains.

The primary objective of this research is to support neurobiologists in understanding the complex functioning of the brain by providing detailed anatomical studies through high-resolution 3D histological and microscopic volume analyses. A deep learning-driven framework for the automated segmentation of high-resolution 3D histological mouse brain images was developed. Efficient comparison between genetically mutated and normal mouse brains is enabled by this framework, utilizing a private dataset consisting of histological microscopic 3D volumes in nearly raw raster data (nrrd) format, with individual file sizes ranging from 25 to 35 GB.

The deep learning models used in this study include nnU-Net, which automates the configuration of segmentation pipelines, and the Segment Anything Model (SAM), adapted for 3D medical imaging through the MedSAM and MaskSAM frameworks. The computational environment was optimized for large-scale data processing, leveraging advanced neural network architectures and high-performance computing resources.

For binary segmentation between the background and the whole brain tissue, 11 full-brain volumes were considered, achieving a Dice Similarity Coefficient (DSC) of 0.99 ± 0.01, while for multi-class segmentation of 24 brain regions, 14 half-brain volumes were prepared, achieving a global DSC of 0.87 ± 0.01. The segmentation time was significantly reduced by our method from approximately 30 hours per volume to just 5 minutes, thereby accelerating the research process.

Our approach demonstrates high precision and robustness in the segmentation of histological mouse brain image, facilitating further research and innovation in computational neuroscience and biomedical imaging. This efficiency not only enhances the feasibility of large-scale studies but also supports high-throughput data processing in brain histology.

*Keywords:* Automated Segmentation, Histological Imaging, 3D Microscopy, High-Resolution Imaging, Mouse Brain, Computational Neuroscience

## 1. Introduction

Understanding the neurobiological basis of brain function and structure is fundamental in neuroscience (Cisneros et al., 2023). The intricate architecture of the brain, encompassing macroscopic features like regional volumes and shapes, as well as microscopic details such as neuronal organization and connectivity, is shaped by both genetic and environmental factors. This complexity underpins the brain's development, function, and susceptibility to various disorders. Neuroanatomical phenotypes—observable structural traits—provide critical insights into these processes. The use of animal models, particularly mice, has been pivotal due to their genetic similarity to humans and the ease of genetic manipulation. Large consortium like the International Mouse Phenotyping Consortium (IMPC) or Knock Out

Mouse Project (KOMP) have set out to address gene function on unprecedented scales, systematically disrupting genes in isogenic mice to study their effects on multiple systems. In 2019, the NeuroGeMM laboratory at the University of Burgundy, France, is a long term collaborator of several centers from the IMPC and has long standing experience in neuroanatomic histological screens in rodent. Thus, they published the first NAP map after screening +1500 gene knockouts. 198 genes impacting brain morphogenesis were identified and interestingly, many of these genes were unknown brain morphogenes (Collins et al., 2019). More recently, an analysis of 20 out of 30 potential autism-related genes at the 16p11.2 locus highlighted MVP as a critical morphogene and candidate gene for the first time (Kretz et al., 2023). The assessment of anatomical abnormalities was primarily performed on high-resolution, 2D histological images, with delineation of the mouse brain regions achieved through manual contouring or semi-automated software-assisted methods. Recently, the laboratory acquired a system for high-resolution histological imaging of mouse brains (in essence, a serial block face imaging device) allowing for detailed 3D neuroanatomical analysis, revealing precisely how specific genes influence brain development and pathology using Neuroanatomical phenotypes (NAPs) as endophenotypes of neurodevelopmental diseases. Manual segmentation of these images, however, is labor-intensive and time-consuming, necessitating the development of automated methods to enhance efficiency and accuracy in neuroanatomical studies.

Deep learning, a subset of artificial intelligence (AI), has revolutionized the field of medical image analysis through its ability to learn complex patterns and features from large datasets. Convolutional neural networks (CNNs), particularly architectures like U-Net and its variants, have become the gold standard for image segmentation tasks. These models can automatically delineate structures within histological images, significantly reducing the need for manual annotation. In neuroanatomical studies, deep learning models are trained on high-resolution images of brain sections to identify and segment various regions of interest (ROIs). For instance, the NeuroGeMM laboratory, in collaboration with ICMUB (Dijon), harnessed U-Net and Attention U-Net architectures to segment 2D brain histological sections with high precision. Notably, the Attention U-Net incorporates attention mechanisms that dynamically focus on relevant parts of the image, enhancing the model's accuracy. These advancements not only accelerate the annotation process but also maintain high levels of accuracy, as demonstrated by Dice Similarity Coefficients (DSC) exceeding 90% for most brain regions. By integrating deep learning into neuroanatomical research, scientists can handle large-scale phenotyping tasks more efficiently, facilitating the discovery of gene-function relationships and the development of new insights into brain development and disorders (Cisneros et al., 2023).

Now that the laboratory is acquiring 3D volumes, the need for deep learning methods to automatically segment brain regions has become not just a luxury but an essential requirement. Hence, the primary objective of this study is to enhance rodent brain segmentation through the optimization of U-Net-based architectures. Our work aimed at implementing these architectures and developing a pipeline designed to significantly reduce the time required by neuroanatomists for segmentation tasks. Our main strategy was to use the foundational U-Net model, initially proposed by (Ronneberger et al., 2015) but use the most recent development nnU-Net, short for "no-new-Net," which is a self-configuring framework for neural network-based biomedical image segmentation (Isensee et al., 2021). nnU-Net automatically adapts its architecture, preprocessing, and training strategy to suit the specific characteristics of the dataset it processes. This adaptability allows nnU-Net to achieve state-of-the-art performance across a variety of segmentation tasks without manual tuning of its parameters. The method's robustness and effectiveness have significantly influenced medical image analysis, demonstrating the power of adaptive, data-driven approaches in deep learning applications in healthcare.

Several critical areas were addressed through key contributions. High-resolution histological 3D volumes, with data sizes ranging from 25 to 35 GB, were managed effectively. Image processing techniques, including upscaling, downscaling, re-sampling, and curve approximation, along with conversions between nrrd and nifti formats, were implemented to facilitate the initial and final stages of the training pipeline. The nnU-net architectures were employed for model training and testing to enhance the segmentation process. Additionally, a user-friendly tool for automatic segmentation of histological mouse brain images across 24 regions of interest was developed and deployed, significantly outpacing traditional human annotation speeds.

## 2. State of the art

Recent advancements in biomedical image segmentation, fueled by continuous research and innovation, aim to enhance accuracy, efficiency, and applicability across diverse medical imaging modalities and uses. These improvements empower physicians to make better-informed decisions, elevate patient care, and expedite medical research and diagnostics. This section will explore segmentation models, beginning with manual approaches and progressing to fully automated methods, specifically focusing on the segmentation of histological images of the mouse brain.

### 2.1. Manual segmentation

Accurate morphometric description of neurological disorders in human patients is a field that can only progress with the recent development of AI-assisted image segmentation. Currently, simple metrics are often used in the clinic but better description of phenotype would largely benefit disease detection, comprehension and classification. The starting point lays in having sufficient resolution to see structures followed by manual segmentation. Many softwares in preclinical and clinical settings exist and all share the general principle of allowing manual segmentation. An example specifically designed for neuroimaging is brainvoyager (BrainVoyager, 2023). To elaborate the ground truth however, the laboratory used Slicer3D to draw, edit, and manage segmented regions in 3D images. Whilst these software are not approved for clinical work, they are used around the world in preclinical studies. These resources are crucial for researchers and clinicians who require precise segmentation capabilities for detailed brain imaging studies but manual segmentation requires a significant amount of time.

### 2.2. Semi-automatic segmentation

Manual segmentation, while highly accurate, is impractical for large 3D datasets due to the extensive time and effort required. On the other hand, fully automatic segmentation methods, despite their efficiency, often fail to achieve the necessary accuracy, necessitating user intervention for corrections which are even more time-consuming in 3D.

Jones et al. introduced a method for semi-automatic segmentation where it is assumed that each $I_i$ is sufficiently oversegmented into $r$ superpixels, $o_j$, such that each true region, $t_l$, can be generated from $o_j$ as in Equation 1. Using this assumption, the method reorganizes the initial segmentation, $P_I$, predicted by the automatic method by utilizing the hierarchical structure, each superpixel $o_j$, and user input to generate the final predicted segmentation, $P_F$. If the results are ideal, then $P_F = T$. Due to the 2D nature of the automatic segmentation method used, both proofreading the 2D segmentation and linking the resulting 2D segments using automated suggestions are necessary. By completing both of these steps simultaneously, the method aims to reduce the amount of time required from the user.

$$t_l = \bigcup_{j \in \gamma_l^T} o_j \tag{1}$$

Liangjia Zhu et al. present an advanced method for segmenting anatomical structures in medical imagery. This research, conducted across Stony Brook University, the University of Alabama at Birmingham, and Harvard Medical School, reformulates the GrowCut algorithm as a clustering problem and effectively allows to complete segmentation using partial information such as to segment only two sections 10 to 20 sections a part and letting the algorithm complete the missing segmentation (Zhu et al., 2014). The proposed method leverages the Dijkstra algorithm to enhance computational efficiency, allowing for real-time interaction and application to high-resolution images. The paper highlights the method's efficiency and accuracy through extensive testing on challenging datasets, demonstrating its potential for integration into medical imaging platforms like 3D Slicer.

Jones et al. proposed method integrates automatic segmentation with user-guided corrections to streamline the proofreading process of neuron tracing in EM images (Jones et al., 2015). The automatic component generates a hierarchical structure of superpixels, recommending potential merges that are then reviewed by the user. This hierarchical approach allows users to quickly identify and correct segmentation errors, significantly improving the accuracy of the final segmentation results. The method's efficacy is demonstrated through tests on multiple datasets, showing that even novice users can achieve accuracy levels comparable to expert manual segmentation, but with substantial time savings.

Uberti et al. presents a novel method for brain extraction from MRI data in mice. This technique leverages a level-set method with user-defined constraints to enhance accuracy (Uberti et al., 2009). The authors highlight the limitations of existing methods when applied to mouse brain MRI, especially in low contrast conditions, and propose the Constraint Level Sets (CLS) method as a solution. The CLS method integrates anatomical knowledge to improve the extraction process. They compared the development of both 2D and 3D implementations of this technique and compares their performance using high-resolution T1-weighted (T1-wt) FLASH and T2-weighted (T2-wt) RARE MRI data. The results demonstrate that CLS outperforms traditional seed-based region growing (SBRG) methods, particularly in scenarios with minimal contrast between brain and non-brain tissues. Key findings include the 2D implementation of CLS being slightly more efficient than the 3D version, with both providing significant improvements over SBRG, the accuracy of brain extraction being higher in T2-wt RARE MRI compared to T1-wt FLASH MRI due to better contrast, and the CLS method consistently yielding high overlap measures (OM) with manual segmentation, indicating reliable performance. This technique is applicable to a variety of MRI scans and can be extended to segment other organs and tissues, making it a valuable tool in preclinical neuroscience research using mouse models of neurodegenerative diseases.

### 2.3. Automatic segmentation

Scheenstra et al., presents a novel, efficient approach for the segmentation of various structures in mouse

brain MRIs, both in vivo and ex vivo. The study addresses the challenges posed by low signal-to-noise ratios and low contrast between structures in mouse brain images. The proposed method involves an initial rough affine registration to a template followed by a clustering algorithm that refines the segmentation near the edges. Compared to manual segmentations, the method achieved an average kappa index of 0.7 for 7 out of 12 structures in vivo MRIs and 11 out of 12 structures in ex vivo MRIs. Notably, the method is eight times faster than traditional nonlinear segmentation methods. This automatic segmentation technique is effective for image registration, volume quantification, and annotation of brain structures, thus offering a significant improvement in processing time and accuracy for neuroanatomic studies (Scheenstra et al., 2009).

Tappan et al. introduce NeuroInfo, a novel brain navigation system designed to automate the identification and delineation of brain regions in histologic mouse brain sections. This system functions similarly to a GPS in a car, by registering digital images of experimental mouse brain sections with a three-dimensional (3D) digital mouse brain atlas based on the Allen Mouse Brain Common Coordinate Framework (CCF v3). NeuroInfo retrieves graphical region delineations and annotations from the 3D atlas and superimposes this information onto the digital images of the brain sections, facilitating accurate identification of brain regions without observer bias. Validation studies demonstrated that NeuroInfo performs exceptionally well in delineating large or dorsally located regions, irrespective of the imaging modality used (fluorescence or bright-field microscopy). The implementation of NeuroInfo thus offers a significant advancement in the systematic analysis of brain sections, improving the efficiency and accuracy of brain region identification in neurogenomics, transcriptomics, proteomics, and connectomics studies (Tappan et al., 2019).

Cisneros et al. introduce 2D histological segmentation, it is a crucial technique in neuroanatomical research, providing detailed insights into the structure and organization of brain tissues(Cisneros et al., 2023). This process involves the precise delineation of different anatomical regions within high-resolution histological images, which are typically obtained from thin sections of brain tissue stained to highlight various cellular components. Accurate segmentation of these images is essential for quantifying morphological features, identifying structural abnormalities, and linking these observations to genetic and environmental factors.

Deep learning has emerged as a powerful tool for 2D histological segmentation, offering the ability to automate and standardize the process. Convolutional neural networks (CNNs), such as U-Net and its derivatives, have shown great promise in this domain. These networks are designed to learn from annotated datasets, identifying patterns and features that distinguish different anatomical regions. U-Net, for example, uses an encoder-decoder architecture that captures spatial hierarchies in the images, making it well-suited for biomedical image segmentation. The Attention U-Net, an enhanced version, incorporates attention mechanisms that allow the model to focus on the most relevant parts of the image, thereby improving segmentation accuracy (Segmentation pipeline Figure 1).

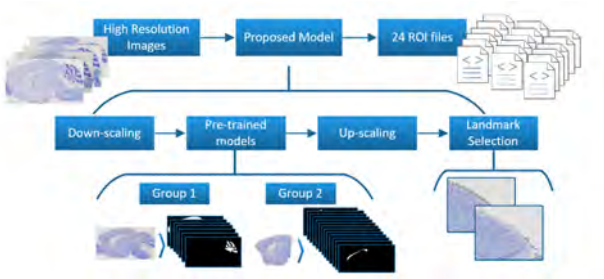



Figure 1: Attention U-Net 2D histological segmentation pipeline (Cisneros et al., 2023)

By integrating automated 2D histological segmentation into neuroanatomical research, scientists can accelerate the analysis of brain structures, enabling large-scale studies that were previously impractical. This not only enhances our understanding of brain development and disorders but also facilitates the discovery of new genetic and molecular targets for therapeutic intervention. As deep learning techniques continue to evolve, their application in histological segmentation is expected to become even more robust, offering greater accuracy and broader applicability in the field of neuroscience (Cisneros et al., 2023).

The nnU-Net methodology is designed to automate the configuration of deep learning-based segmentation pipelines for biomedical imaging tasks (Isensee et al., 2021). The method involves several key components: Initially, nnU-Net processes the provided training data by cropping the images to their non-zero regions, which improves computational efficiency. It creates a dataset fingerprint capturing image size, image spacing, modalities, number of classes, and intensity values, computed over all training cases. nnU-Net then generates a pipeline fingerprint that condenses domain knowledge into heuristic rules. These rules operate on the dataset fingerprint and project-specific hardware constraints to infer necessary design choices. nnU-Net uses a fixed architecture template closely following the original U-Net and its 3D counterpart. Adjustments include the use of instance normalization and leaky ReLUs, and the inclusion of deep supervision for training stabilization. The method involves initializing patch size based on the median image shape after resampling and iteratively adapting network topology, including the number and position of pooling operations, feature map sizes, and convolutional kernel sizes, to fit within GPU memory constraints while maintaining a minimum

batch size of two. nnU-Net automatically generates three U-Net configurations (2D U-Net, 3D U-Net, and 3D U-Net cascade) and uses cross-validation to choose the best-performing configuration or ensemble. Empirical post-processing steps, such as non-largest component suppression, are applied if they improve performance. Training is performed with fixed parameters like learning rate, optimizer, and data augmentation strategies. Inference is conducted using a sliding window approach with Gaussian patch center weighting for smoother predictions. nnU-Net's automated configuration has been tested across diverse datasets, showing strong generalization capabilities. It handles various imaging modalities and target structures effectively, often outperforming specialized pipelines tailored for specific tasks (Isensee et al., 2021).

## 3. Material and methods

### *3.1. Computational environment*

The system runs CentOS Linux 7.7 "Core" edition, tailored for enterprise deployments to provide a stable and secure environment for production setups. It operates on the Linux kernel, optimized for performance on x86_64 architectures. The CPU is an Intel(R) Xeon(R) Gold 6226 @ 2.70GHz, capable of boosting up to 3.70 GHz, and supports 64-bit operations. It features a dual-socket setup with 24 cores per socket, totaling 48 cores, and is hyper-threaded, offering 96 threads. The CPU has a BogoMIPS rating of 5400.00, supports virtualization via Intel VT-x, and includes a cache layout with 32 KB L1 cache for data and instructions per core, 1 MB L2 cache per core, and a 19.7 MB L3 cache shared across the CPU.

### *3.2. Dataset*

Our private dataset consists of histological microscopic 3D volumes of mouse brains. Each volume is provided in a nearly raw raster data (nrrd) format, with individual file sizes ranging between 25 to 35 GB. For binary segmentation between the background and the whole brain tissue, 11 full brain volumes have been prepared, as illustrated in Figure 2 (a). Out of these, 9 volumes are designated for training purposes, while the remaining 2 volumes are set aside for validation.

Additionally, for the segmentation of specific brain regions, we utilized 14 half-brain volumes, as depicted in Figure 2 (c). Among these, 11 volumes are allocated for training, and 3 volumes are used for validation.

This dataset is designed to facilitate advanced research and development in the field of brain histology and segmentation and remains private for controlled access and use.

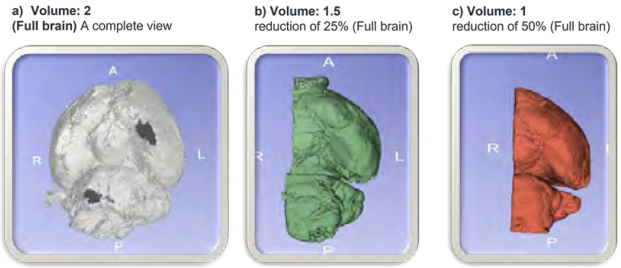



Figure 2: 3D View of mouse brain volume sequence: (a) Volume 2: Full brain complete view, (b) Volume 1.5: 25% reduction of full brain, (c) Volume 1: 50% reduction of full brain.

| SL | TAG | NAME |
|---|---|---|
| 1 | CTX+ | Cortex |
| 2 | cc+ | Corpus callosum |
| 3 | CPu | Caudate Putamen |
| 4 | DG | Dentate Gyrus |
| 5 | HP | Hippocampus |
| 6 | RHP | Retro hippocampus |
| 7 | A | Amygdala |
| 8 | ig | Indusium griseum |
| 9 | fi | Fimbria |
| 10 | f | Fornix |
| 11 | st | Stria terminalis |
| 12 | ic | Internal capsule |
| 13 | och | Optic chiasm |
| 14 | ac | Anterior commissure |
| 15 | fr | Fornix |
| 16 | Hb | Habenula |
| 17 | TH | Thalamus |
| 18 | HY | Hypothalamus |
| 19 | MB | Midbrain |
| 20 | P | Pons |
| 21 | MY | Medulla |
| 22 | TCB | Total Cerebellum Brain |
| 23 | V | Ventricle |
| 24 | OB | Olfactory bulb |

Table 1: Neuroanatomical Features of 24 Regions in the Mouse Brain.

### *3.3. Dataset Preparation and GT preparation*

The dataset consists of a 3D volume and its corresponding ground truth (GT) data Figure 4.

The processing of 3D imaging data involves several critical steps for the preparation of the data to ensure accuracy and efficiency across various applications. Initially, downsampling the image and GT Figure 8 (b) by a factor of 5 is essential. This reduction in resolution helps in decreasing the computational load and storage requirements, while care is taken to preserve vital details necessary for accurate analysis.

Subsequent to downsampling, converting file formats, such as from `.nrrd` to `.nifti`, ensures compatibility across different software tools. This step is crucial for seamless integration and manipulation of data within

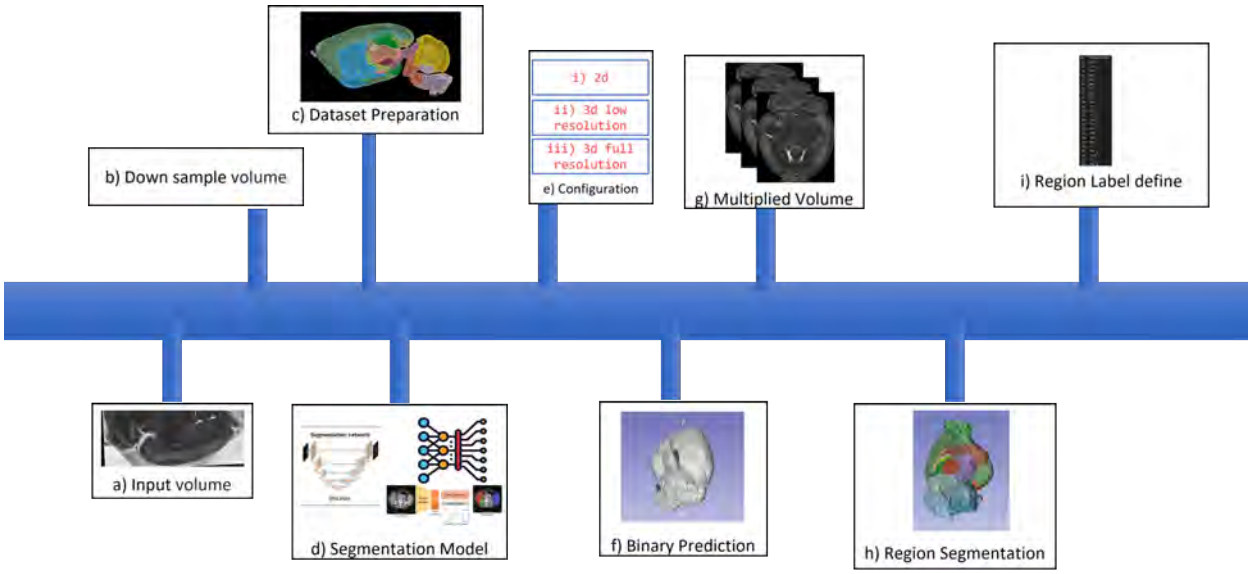



Figure 3: Full pipeline of the segmentation task. (a) Input 3D volume. (b) Downsample volume. (c) Dataset preparation. (d) Segmentation model. (e) Configuration: Specification of image type and resolution, choosing between 2D, 3D low resolution, and 3D full resolution. (f) Binary prediction (g) Multiplied volume. (h) Region Segmentation (i) Region labels.

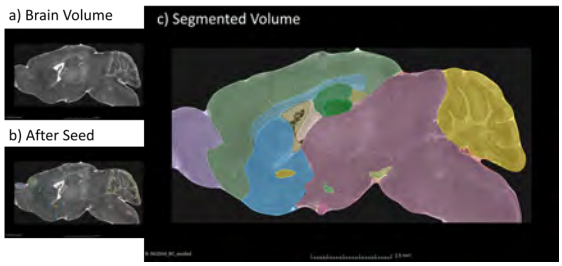



Figure 4: Segmented volume visualization, GT seed to segmentation: (a) Histological brain volume, (b) Seeded regions, (c) Segmented volume grown from seeds.

various processing environments.

Another important task is the reshaping of the volume and GT Figure 3 (c). This process adjusts the dimensions of the data to fit specific tools for analysis or visualization, maintaining the original aspect ratios to prevent distortion of spatial relationships in the data.

Following reshaping, an intensity check on both the volume and GT is conducted. This involves examining the pixel intensity distributions to ensure they meet predefined criteria necessary for further processing. This check helps in identifying any anomalies that might affect subsequent analyses, such as segmentation or classification.

The segment label check on GT ensures the accuracy and consistency of segmentation labels. This step involves verifying that each label correctly identifies the corresponding segment and checking for errors such as mislabeling or overlapping labels, which are crucial for the reliability of downstream analyses.

In Slicer3D 3D Slicer, the seed growing technique for segmentation is applied Figure 5. This method starts with manually or automatically selected seed points within the image. These points serve as the basis for the segmentation algorithm, which then expands to adjacent areas based on predefined criteria such as intensity thresholds or color similarities. This technique is especially effective for isolating specific structures and can be tailored to specific needs by adjusting the growth criteria.

Each of these steps is integral to the preprocessing workflow, setting a strong foundation for robust and precise analyses of 3D imaging data.

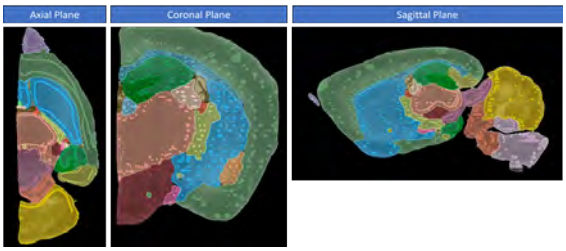



Figure 5: Half brain segmentation using grow from seeds: 2D view without background, Displaying Axial, Coronal, and Sagittal Planes.

The GrowCut algorithm (in the grow from seeds module of Slicer3D) by Liangjia Zhu et al. utilizes a small number of user-labeled pixels to guide the segmentation process, which is iteratively refined by a Cellular Automaton. The method allows users to observe and interact with the segmentation evolution, making adjustments in challenging areas while leaving reliably segmented regions untouched.

Key contributions and features of the GrowCut method include its effectiveness in handling moderately

complex segmentation tasks and its applicability to images of any dimension ($N \geq 1$). The method is efficient in performing multi-label segmentation without increased computation time due to the number of labels, which is crucial for large-scale or detailed tasks. Additionally, GrowCut is highly extensible, allowing the creation of new segmentation algorithms with specific properties tailored to various needs. Its high interactivity is a standout feature, enabling users to refine segmentation continuously during the process, thus enhancing precision and user control. The GrowCut algorithm has been tested on both generic photographs and medical images, demonstrating that it requires only modest user effort to segment moderately difficult images effectively. Vezhnevets et al, highlight the ongoing relevance of semi-automatic segmentation methods, given the limitations of fully automated techniques in providing guaranteed results across diverse scenarios. This blend of efficiency, flexibility, and user interactivity positions GrowCut as a valuable tool in the segmentation landscape. Vezhnevets and Konouchine provides

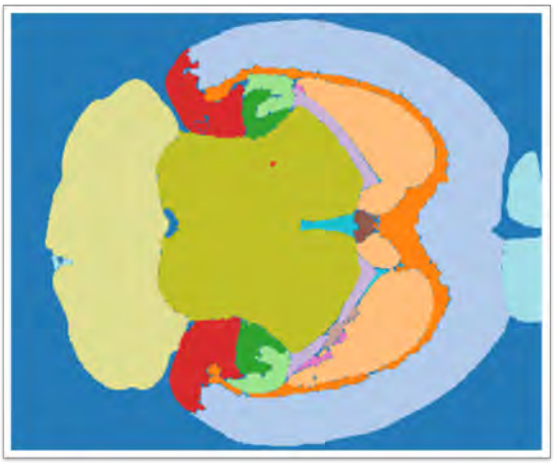


Figure 6: GT of 2D View: Full brain segmented by individual regions.

a review of related interactive segmentation techniques, such as graph cuts, random walker, and region growing methods, and positions GrowCut as a competitive alternative, particularly noted for its user convenience and segmentation quality in multi-label tasks Vezhnevets and Konouchine (2005).

Biomedisa is a highly effective "fill between slice" algorithm used extensively in the field of biomedical imaging (Biomedisa). It excels in reconstructing 3D models from 2D image slices by accurately filling the gaps between these slices, significantly enhancing the quality and continuity of 3D reconstructions. This makes Biomedisa an invaluable tool for researchers and professionals dealing with complex biological structures.

### *3.4. Pre-processing*

In histological imaging, preprocessing is essential to enhance image quality and ensure accurate analysis. During preprocessing, we often encounter artifacts such as bubbles and overlay lines, which obscure critical information and compromise data integrity. To address these challenges, we employed a combination of median filtering and morphological operations.

Median filtering, a non-linear digital filtering technique, was chosen for its effectiveness in noise reduction while preserving edge details. Specifically, a 3x3x3 median filter was applied to the images. This filter replaced each pixel's value with the median value from its local neighborhood, effectively reducing noise and smoothing out bubble artifacts. The result was a significant reduction in noise without a substantial loss of image detail.

To further refine the images, morphological operations were employed, specifically the processes of opening and closing. These operations are fundamental in image processing, particularly for removing small artifacts and smoothing object boundaries. Morphological opening, which involves erosion followed by dilation, was used to remove small objects such as bubbles from the foreground of the image. Morphological closing, which involves dilation followed by erosion, was employed to fill small holes and eliminate overlay lines.

Various structuring elements were tested to optimize these operations, including (3,1), (3,2), and (3,3) configurations. The best results were obtained using a (3,3) structuring element, which effectively removed both bubble and overlay line artifacts without distorting the significant features of the images. Additionally, applying these morphological operations to binary images yielded particularly good results for line artifacts.

In conclusion, the preprocessing pipeline combining a 3x3x3 median filter with morphological opening and closing using a (3,3) structuring element proved highly effective in removing artifacts from histological images. This approach ensured high-quality, artifact-free images, facilitating more accurate and reliable scientific analysis.

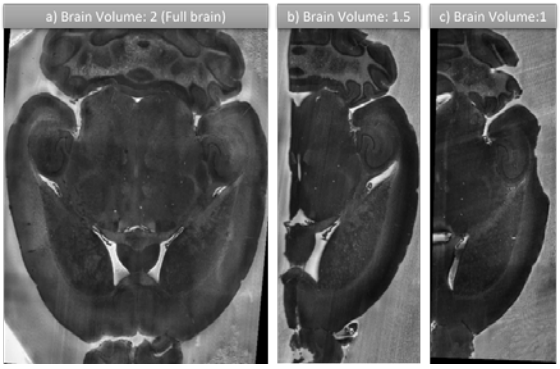



Figure 7: 2D View of Mouse Brain Volume Sequence. Figure 2

### *3.5. Data preparation*

After preparing the database and completing the review stages, region masks for each region of interest within the mouse brain were obtained (Figure 6). The variation in the number of masks is due to factors like information loss, and mislabeling in the brain region. Once the binary image dataset was created, the deep learning training stage was initiated.

### *3.6. Deep Learning Models*

The Complete Pipeline of the Segmentation is presented in Figure 3, The segmentation task begins with acquiring a 3D histological mouse brain volumetric dataset (a). This high-resolution dataset is then downsampled (b) to reduce its resolution, facilitating easier handling and processing. The prepared dataset (c) is organized and formatted to ensure compatibility with the subsequent steps. A segmentation model is applied (d) to delineate different regions within the 3D volume. The configuration step (e) involves specifying the image type and resolution, with options including 2D, 3D low resolution, and 3D full resolution. The model generates binary predictions (f) for the full brain (see Figure 2(a)) and half brain (see Figure 2(c)). The input volume is then multiplied by these binary predictions (g) to obtain segmented volumes for both the full brain and half brain. These segmented regions are detailed in Table 1. (h). Finally, region labels are defined within the 3D volume as metadata, providing a comprehensive description of each region as outlined in Table 1.

**nnUNet: a self-configuring method** In our study, the nnUNet configuration for medical image segmentation has been carefully designed to meet different resolution needs, ensuring efficient and accurate analysis.

For the 2D configuration, the batch size was set to 6, allowing for effective memory usage and stable training Figure 3 (e.i). The chosen patch size of 896x640 aligns well with the median image size of 782.0x633.0 voxels, ensuring that most of the image content is included within each patch, thereby maximizing the utility of the data without excessive padding or cropping.

For the 3D low-resolution configuration, a batch size of 2 is utilized, balancing the memory requirements and processing efficiency for 3D image volumes Figure 3 (e.ii). The patch size of 80x192x160 covers substantial portions of the images, which typically measure around 135x295x239 voxels. The spacing of 2.65mm in all dimensions reduces the computational load, allowing for faster training iterations while capturing broad anatomical structures at a coarser level.

In the 3D full-resolution configuration, the batch size is maintained at 2 to handle the high memory demand of detailed 3D data Figure 3 (e.iii). The patch size remains consistent with the low-resolution setting at 80x192x160, facilitating a uniform training approach. This configuration is tailored for higher resolution images, with a median size of approximately 359.0x782.0x633.0 voxels and a spacing of 1.0mm in all dimensions, capturing fine anatomical details crucial for precise segmentation.

**Background removal:** Background and Whole-Brain segmentation, A robust preprocessing methodology was employed for binary segmentation and whole-brain extraction using the nnUNet framework. The dataset comprised volumetric brain images formatted for nnUNet compatibility. Necessary environment variables were configured to ensure smooth operation, setting paths for the dataset and model configurations. The dataset consisted of 14 brain volumes, split into 11 for training and 3 for validation, providing a comprehensive training set while reserving sufficient data for model evaluation.

The segmentation pipeline began with the input of large 3D brain images, each with dimensions around (4128, 2978, 1844). These images were downsampled by a factor of five to approximately (654, 783, 383) to reduce computational load and memory usage, maintaining adequate detail for accurate segmentation. The nnUNet model was utilized for its adaptive capabilities and superior performance in medical image segmentation. Three configurations were implemented: a 2D configuration that processed slices independently, a 3D low-resolution (3D low resolution) configuration, and a 3D full-resolution (3D full resolution) configuration.

The trained nnUNet model generated binary masks of the 3D volume, The training was conducted using five-fold cross-validation to ensure robustness. Each fold involved training on a subset and validating the remaining data. The resulting binary masks were then multiplied by the downsampled images to extract the segmented brain regions.

**Hippocampus segmentation** is an important structure to analyze in given its role in cognitive functions, especially memory formation and spatial navigation.

The hippocampus segmentation pipeline begins with a high-resolution input volume of dimensions (4128, 2978, 1844). To manage computational demands and memory constraints, this 3D volume is downsampled by a factor of 5, reducing the dimensions to approximately (826, 596, 369). The nnUNet model, is then employed in three configurations: 2D, 3D low-resolution, and 3D full-resolution. Each configuration leverages the strengths of both 2D and 3D convolutions, balancing computational efficiency and spatial context. The model predicts a binary mask 3D volume, identifying the hippocampus (foreground) from the rest of the brain (background). This binary mask is then upsampled back to the original high-resolution dimensions, ensuring precise alignment with the initial input volume. This pipeline efficiently processes large volumes while maintaining high segmentation accuracy, leveraging nnUNet's adaptive configuration capabilities.

**Segmentation of 24 Regions** including background 25 regions (Figure 13) The multiclass segmentation

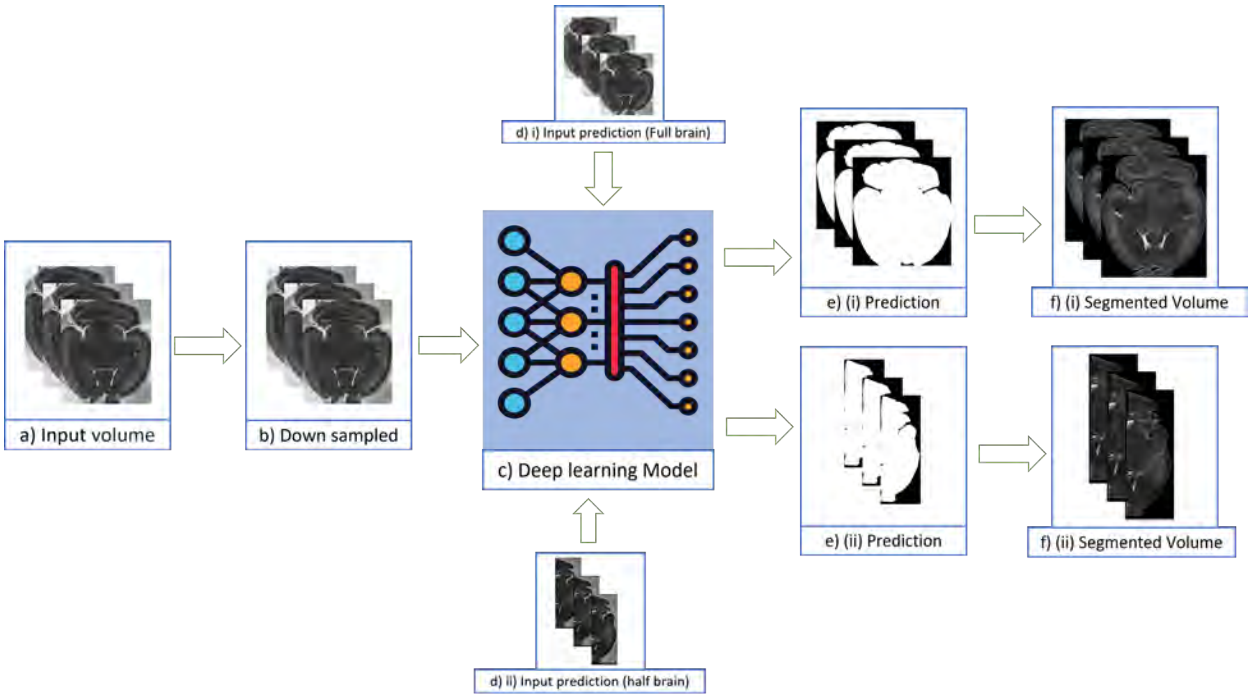



Figure 8: Binary Segmentation Pipeline: (a) Input Volume Dimensions: (4128, 2978, 1844), (b) Downsampled 3D Volume by a Factor of 5, (c) Deep Learning Model: nnUNet, (d) Prediction Inputs: (i) Full Brain Volume, (ii) Half Brain Volume, (e) Prediction Outputs: Binary Mask 3D Volume for (i) Full Brain, (ii) Half Brain, (f) Final Results: Segmentation of (i) Full Brain (combining d(i) with e(i)), (ii) Half Brain (combining d(ii) with e(ii)).

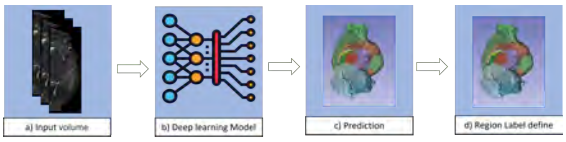



Figure 9: Half Brain Segmentation Pipeline: (a) Input volume, (b) Deep Learning Model, (c) 24 Region Prediction Volume, (d) Region Label Definition. The pipeline includes the stages from initial input of the brain volume through deep learning processing, resulting in the prediction of 24 segmented regions, and the subsequent definition and labeling of these regions.

pipeline begins with an original high-resolution image (Figure 8b) that is downsampled to reduce its dimensions for efficient processing. This downsampled image is then combined with a binary prediction mask (Figure 8) generated by a deep learning model. The regions of interest are isolated by multiplying the downsampled image with the binary mask. The resulting combined image is subsequently used as input for region segmentation, where advanced segmentation algorithms classify each pixel or voxel into multiple classes, representing different anatomical structures or regions of interest.

During the training phase of the multiclass segmentation pipeline, the process is divided into multiple folds to ensure robust model performance and to prevent overfitting. Specifically, for the 3D low-resolution configuration, the training is conducted across five folds, labeled from 0 to 4. Each fold represents a distinct subset of the data used for training and validation. The training times for each fold took 30, 33, 30, 20 and 20 hours, respectively for folds ”0” to ”4”. This distributed training approach not only enhances the model’s ability to generalize but also provides a comprehensive evaluation of its performance across different subsets of the dataset. The varying training times reflect the computational effort and complexity associated with each fold, ensuring that the model is thoroughly trained and validated.

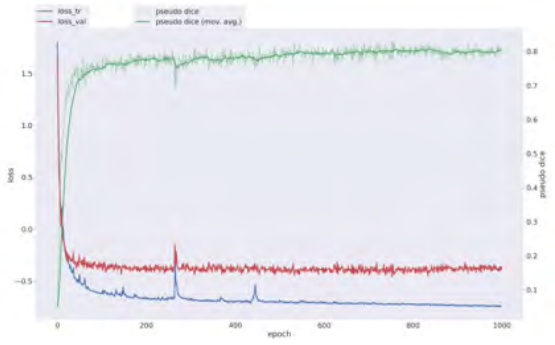



Figure 10: Training and validation performance over epochs. The graph shows the training loss (loss_tr) and validation loss (loss_val) decreasing steadily, indicating the model’s learning process. Additionally, the pseudo Dice coefficient (pseudo dice). Pseudo Dice scores are particularly useful during training to monitor progress and performance on specific patches, helping to identify areas where the model may be underperforming and requires more focus Jubair and R. (2023).

During the training phase for the 3D full-resolution configuration of the multiclass segmentation pipeline, the process is also divided into five folds, labeled from 0 to 4. The training duration for each fold is as follows: fold 0 takes 36 hours, while folds 1, 2, 3, and 4 each take 35 hours. This approach ensures thorough training and validation, leveraging the higher-resolution data to achieve more precise segmentation results. The slightly longer training duration for fold 0 indicates the initial computational effort required, while the consistent 35-hour training times for the subsequent folds reflect a stable training process. This comprehensive training across all folds ensures that the model is robust and capable of generalizing well across different subsets of the dataset.

In the final step of the multiclass segmentation pipeline, after segmenting the regions, the metadata is updated to include region names. This involves assigning meaningful labels to each segmented region, ensuring that the segmentation results are not only accurate but also interpretable. Each region, represented by specific intensity values, is given an anatomical name to reflect its identity. For example, intensity value 5 might correspond to the "hippocampus," while intensity value 23 might denote the "ventricle." Table 1 By renaming each region in the metadata according to its intensity value, the segmentation output becomes more informative and useful for further analysis and clinical applications. This integration of region names with intensity-based metadata enhances the overall utility of the segmentation pipeline, providing a clear and contextually relevant map of the segmented regions, facilitating better understanding and communication of the results.

### *3.7. Deep Learning Framework*

The segmentation task workflow involves several critical steps to accurately process and analyze 3D volumetric data. The process begins with the input of a 3D volume, which is a detailed dataset representing the structure to be analyzed. Following this, the volume undergoes downsampling to reduce its resolution, facilitating easier handling and processing without significant loss of critical information. The dataset preparation phase follows, where the input data is organized and formatted appropriately for the subsequent steps.

Once the dataset is ready, it is fed into a segmentation model, a specialized algorithm designed to delineate different regions within the 3D volume based on the training it has received. This model's configuration is critical and involves specifying the image type and resolution to be used: whether a 2D slice-by-slice approach, a 3D low-resolution volume, or a 3D full-resolution volume, depending on the specific requirements and constraints of the analysis.

The segmentation model then produces a binary prediction for the entire brain (Figure 2 (a)) and for half of the brain (Figure 2(c)), effectively differentiating between the regions of interest and the background. This prediction is then multiplied with the input volume, resulting in segmented volumes that visually and numerically represent the areas identified by the model. These segmented volumes are again illustrated for the full brain in Figure 2(a) and for half the brain in Figure 2(c).

The segmented regions are further detailed in 1, which presents the region segmentation results, quantifying the areas identified in the predictions. Additionally, region labels are defined within the 3D volume as metadata, providing a comprehensive reference for each segmented part, also tabulated in Table 1. This meticulous process ensures that each segment is correctly labeled and can be used for further analysis or comparison, maintaining the integrity and accuracy of the segmentation task.

### *3.8. Evaluation Metrics*

#### *3.8.1. The Dice coefficient (DSC)*

The Dice coefficient, or Dice similarity coefficient (DSC), 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:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|} \tag{2}$$

#### *3.8.2. Hausdorff Distance (HD)*

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:

$$HD(A, B) = \max\left(\max_{a \in A}(d(a, B)), \max_{b \in B}(d(b, A))\right) \tag{3}$$

where $d(a, B)$ represents the minimum distance between a point $a$ in set $A$ and the closest point in set $B$, and $d(b, A)$ represents the minimum distance between a point $b$ in set $B$ and the closest point in set $A$.

#### *3.8.3. Relative Absolute Volume Difference (RAVD)*

Relative Absolute Volume Difference (RAVD) is a crucial metric in scientific research for evaluating the accuracy and reliability of volumetric measurements, particularly in medical imaging, geospatial studies, and 3D modeling. Defined as

$$\text{RAVD} = \frac{|V_{\text{ref}} - V_{\text{target}}|}{V_{\text{ref}}} \times 100\% \tag{4}$$

where $V_{\text{ref}}$ is the reference volume and $V_{\text{target}}$ is the target volume, RAVD provides a clear measure of discrepancy. It is widely used to assess the precision of automated segmentation algorithms in medical imaging,

the accuracy of volumetric measurements in geospatial applications, and the fidelity of 3D reconstructions. A lower RAVD indicates higher accuracy, making it essential for comparative analysis, accuracy assessment, and quality control. However, its effectiveness depends on the reliability of the reference volume and can be influenced by the size of the volumes being compared.

## 4. Results

In this section, the efficiency of our approaches to automatically segment various components is presented in order of difficulty. Firstly, background removal is addressed through background and whole-brain segmentation. Secondly, the segmentation of a single region, specifically the hippocampus, is demonstrated. Lastly, the segmentation of 24 regions in histological mouse brain images is covered.

### *4.1. Background removal*

Binary segmentation background and whole-brain tissue segmentation, utilizing a 5-fold cross-validation approach. For the 2D and 3D low-resolution configurations, training was conducted up to 1000 epochs, while for improved results and experiments in 3D full-resolution, training was extended to 2000 epochs. The 2D configuration yielded the mean best DSC of 99.48%, with training times ranging from 20 to 55 hours. The 3D low-resolution configuration achieved the mean best Dice scores of 99.37%, with training times between 13 and 26 hours. The 3D full-resolution configuration exhibited the mean best Dice scores of 97.82%, with longer training times ranging from 47 to 79 hours. Our findings indicate that the 2D and 3D low-resolution models generally achieved higher Dice scores and required shorter training times compared to the 3D full-resolution model. The comparative analysis between manual and deep learning-based (nnUNet) segmentation results of a full brain 3D volume is Figure 11.The grayscale image on the left serves as a slice from the original 3D volume used for both segmentation processes, highlighting various anatomical structures. The volume measurements obtained from manual segmentation and deep learning segmentation using nnUNet, show excellent agreement between the two methods. The manual segmentation volume is equal to 247.54 mm³ whilst the deep learning segmentation volume is 247.43 mm³ using nnUNet. This comparison demonstrates the high accuracy of the nnUNet model, which closely matches manual segmentation in both volume and visual representation. The deep learning model achieves superior smoothing performance, without loosing important details such as vessels at the surface, underscoring its potential for the precise and reliable tasks of isolating the brain from the background, a prerequisite step to improve performance of sub-volume segmentation but also reducing the physical size of the volumes for subsequent training tasks.

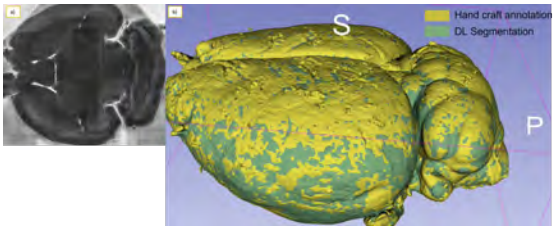



Figure 11: Comparison of manual and deep learning (nnUNet) segmentation results. (a) A cross-sectional 2D view of the brain used for segmentation. (b) A 3D rendering of the brain with segmented regions.

### *4.2. Hippocampus segmentation*

The single region segmentation has been developed in our pipeline. This approach is considered crucial due to the anticipated need for detailed analysis of individual brain regions in the near future. A best DSC score of 0.9411 in 3D low-resolution imaging for hippocampus segmentation has been achieved to ensure robustness and effectiveness.

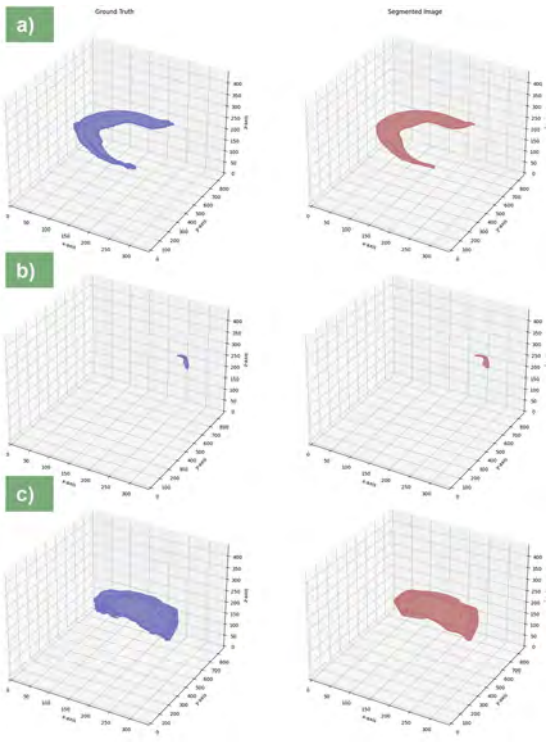



Figure 12: 3D Segmentation Comparison for Different Brain Regions. The ground truth (left) and the segmented image (right). (a) Hippocampus (b) Fimbria: Segment smaller regions accurately. (c) Hypothalamus.

### *4.3. Segmentation of 24 Regions*

The segmentation of 24 regions in histological mouse brain images was performed using a 3D full-resolution

deep learning model with five-fold cross-validation (Figure 13). The model was trained for 1000 epochs, with each fold requiring approximately 35 to 40 hours. This approach achieved precise segmentation results, with the best DSC scores obtained being 0.87. Visual representations of the segmentation results offer detailed comparisons between the ground truth and the model's segmented images for intricate mouse brain anatomical structures like the hippocampus, fimbria (Figure 12 (b)), and hypothalamus. These side-by-side 3D visualizations reveal the model's accuracy in capturing fine anatomical details.

Small region identification of histological mouse brain (Table 2) categorizes the data into five quantile-based categories for better analysis: 1. Micro: 0.66k to 70.99k, 2. Small: 70.99k to 379.68k, 3. Medium: 379.68k to 1,298.83k, 4. Big: 1,298.83k to 2,440.85k, 5. Large: 2,440.85k to 98,251.43k. Rows 8 to 17 of the table represent each region with a relatively small size, identified by pixel counts. The categorization into micro, small, medium, big, and large regions allows for a more structured analysis of the histological mouse brain data. Identifying these small regions is crucial for understanding the detailed anatomical and functional organization of the brain. Smaller regions, in particular, can highlight subtle differences in tissue composition and structure that may be overlooked in broader analyses.

For a visual representation of these categories and their corresponding pixel counts in Figure 23.

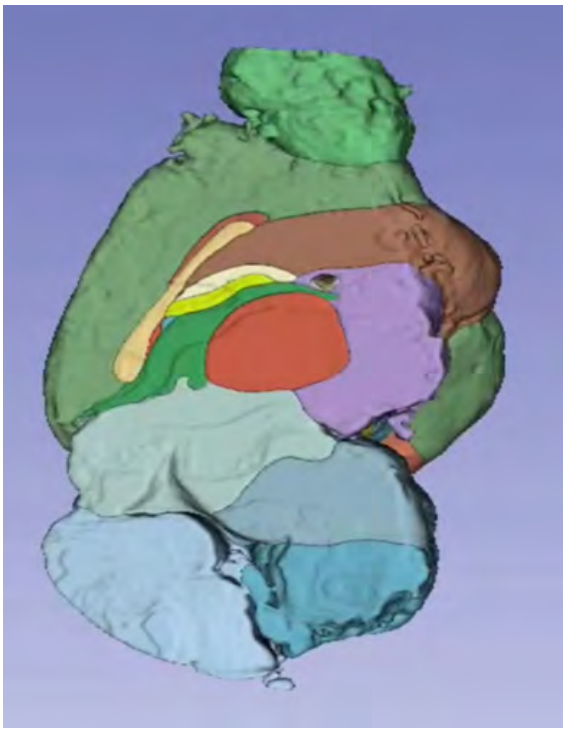


Figure 13: Segmented half brain 24 distinct regions Table 1.

The analysis of Dice Similarity Coefficient (DSC) scores for the segmentation of 3D microscopic volumes of mouse brains reveals varying levels of accuracy across different regions. High-performing regions such as the Cortex (CTX+), Caudate Putamen (CPu), Hippocampus (HP), Midbrain (MB), and Total Cerebellum Brain (TCB) demonstrated consistently high accuracy, with scores approaching 1.0. These results indicate robust and reliable segmentation performance in these regions. In contrast, regions like the Amygdala (A), Indusium Griseum (ig), Anterior Commissure (ac), Internal Capsule (ic), and Stria Terminalis (st) exhibited moderate to low accuracy, highlighting significant variability and inconsistency in segmentation. These findings suggest that further refinement of segmentation algorithms is needed for these regions. Particularly, the Amygdala (A) and Indusium Griseum (ig) require substantial improvements to achieve higher consistency and accuracy. The Anterior Commissure (ac) and Internal Capsule (ic) also demand more robust segmentation methods due to their considerable variability. The Stria Terminalis (st) and Medulla (MY), despite some high accuracy scores, showed inconsistencies and outliers, indicating a need for algorithm enhancement.

The Dice Similarity Coefficient (DSC) measures the overlap between segmented regions, with scores closer to 1 indicating better segmentation accuracy (Figure 14). Regions such as the Cortex (CTX+), Caudate Putamen (CPu), Dentate Gyrus (DG), Hippocampus (HP), and Retro Hippocampus (RHP) show high DSC scores, close to 1, with small interquartile ranges (IQRs), indicating high segmentation accuracy. Conversely, regions like the Amygdala (A), Fimbria (fi), Internal Capsule (ic), Stria Terminalis (st), and Fornix (f) have more varied DSC scores with larger IQRs, indicating lower and more variable segmentation accuracy. Outliers in regions such as the Fimbria (fi), Internal Capsule (ic), and Hypothalamus (HY) indicate occasional poor segmentation performance, often due to the smaller size of these regions.

Hausdorff Distance (HD) measures the maximum distance between boundary points (Figure 15), and values vary widely across regions. Regions like the Cortex (CTX+), Corpus Callosum (cc+), Caudate Putamen (CPu), Dentate Gyrus (DG), Hippocampus (HP), and Retro Hippocampus (RHP) exhibit lower Hausdorff distances with smaller IQRs, indicating better boundary matching and consistency. However, regions such as the Amygdala (A), Fimbria (fi), Internal Capsule (ic), and Stria Terminalis (st) exhibit higher and more variable Hausdorff distances, suggesting less accurate boundary matching. Outliers are present in several regions, particularly in the Cortex (CTX+), Corpus Callosum (cc+), Dentate Gyrus (DG), and Hypothalamus (HY), indicating occasional large deviations in boundary matching accuracy.

Relative Absolute Volume Difference (RAVD), the regions exhibit varying RAVD values Figure 16, indi-

Table 2: Identification of a small region of a histological mouse brain, with all numerical formats presented in thousands. Volumetric samples are denoted with the NG prefix, by unique identifiers

| S.No | Region | NG4108 | NG4111 | NG4116 | NG4119 | NG4115 | NG4120 | NG4114 | NG4117 | NG4109 | NG4112 | NG4110 | NG4113 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | BG | 100013.582 | 52176.648 | 94258.667 | 59210.298 | 70221.045 | 77420.685 | 52349.698 | 55401.809 | 61268.942 | 47907.029 | 98251.432 | 82188.426 |
| 2 | CTX+ | 13123.790 | 12031.753 | 13780.858 | 11017.591 | 15015.588 | 10144.349 | 9309.837 | 11340.841 | 15066.101 | 13047.163 | 14989.756 | 13846.545 |
| 3 | cc+ | 908.045 | 988.205 | 999.139 | 728.859 | 1061.451 | 809.458 | 737.164 | 901.475 | 1030.712 | 1103.185 | 1130.853 | 1071.331 |
| 4 | CPu | 5677.092 | 5888.111 | 5242.927 | 4939.545 | 5739.518 | 5731.060 | 5049.717 | 5393.050 | 5701.843 | 5799.002 | 5882.564 | 5793.639 |
| 5 | DG | 686.849 | 721.753 | 597.888 | 536.471 | 671.236 | 795.151 | 597.663 | 590.658 | 671.340 | 700.306 | 652.725 | 713.284 |
| 6 | HP | 1506.707 | 1580.086 | 1460.669 | 1310.951 | 1706.985 | 1362.181 | 1423.286 | 1425.295 | 1611.106 | 1734.336 | 1810.564 | 1818.336 |
| 7 | RHP | 1297.884 | 1208.729 | 1594.038 | 1240.692 | 1722.468 | 1143.885 | 1148.011 | 1382.542 | 1603.376 | 1648.532 | 1654.333 | 1435.451 |
| 8 | A | 250.754 | 182.951 | 284.201 | 189.176 | 300.290 | 172.005 | 0.664 | 224.049 | 238.222 | 274.419 | 274.820 | 282.921 |
| 9 | ig | 10.214 | 9.567 | 11.761 | 13.625 | 13.165 | 7.224 | 5.478 | 10.612 | 10.908 | 8.309 | 13.005 | 10.483 |
| 10 | fi | 272.023 | 289.190 | 268.655 | 168.731 | 279.582 | 234.104 | 250.916 | 240.480 | 282.728 | 288.411 | 281.690 | 304.269 |
| 11 | ac | 109.093 | 93.982 | 71.913 | 65.628 | 77.476 | 149.950 | 71.942 | 118.400 | 110.986 | 72.515 | 136.343 | 80.568 |
| 12 | ic | 365.518 | 80.659 | 80.808 | 48.385 | 407.570 | 470.984 | 67.236 | 361.389 | 414.435 | 60.174 | 406.359 | 72.852 |
| 13 | st | 77.500 | 397.402 | 411.726 | 341.105 | 66.444 | 53.237 | 313.471 | 65.055 | 66.817 | 429.097 | 72.708 | 336.290 |
| 14 | f | 66.535 | 146.450 | 137.898 | 131.135 | 65.727 | 37.453 | 104.768 | 49.852 | 70.986 | 100.821 | 65.740 | 138.812 |
| 15 | och | 80.672 | 79.258 | 60.538 | 44.736 | 95.518 | 77.056 | 50.204 | 91.356 | 99.020 | 101.268 | 86.433 | 70.648 |
| 16 | fr | 19.389 | 18.241 | 18.111 | 19.954 | 19.267 | 19.807 | 22.124 | 15.584 | 22.303 | 23.305 | 16.784 | 25.441 |
| 17 | Hb | 86.286 | 80.434 | 83.132 | 37.666 | 76.835 | 89.292 | 71.064 | 72.663 | 82.581 | 78.683 | 69.167 | 76.541 |
| 18 | TH | 2087.008 | 2115.503 | 1892.110 | 1447.076 | 1945.981 | 1766.101 | 1620.322 | 1728.938 | 1957.220 | 2002.845 | 2098.648 | 2160.638 |
| 19 | HY | 1445.082 | 1474.367 | 1543.971 | 1295.246 | 1422.002 | 1784.889 | 1385.987 | 1411.380 | 1567.157 | 1253.825 | 1632.031 | 1431.069 |
| 20 | MB | 3766.019 | 3812.262 | 3318.945 | 3457.845 | 3460.886 | 3790.082 | 3478.057 | 3424.259 | 3518.657 | 3753.363 | 3649.733 | 3762.247 |
| 21 | P | 2427.230 | 2336.754 | 2262.427 | 2324.468 | 2271.670 | 2440.851 | 2597.559 | 2314.373 | 3791.796 | 2343.732 | 2528.145 | 2522.491 |
| 22 | MY | 3055.420 | 3278.200 | 2907.002 | 2540.666 | 3225.173 | 3193.210 | 2903.537 | 2630.564 | 7.674 | 3417.776 | 2991.147 | 3246.510 |
| 23 | TCB | 5527.757 | 5066.770 | 4999.864 | 5434.437 | 5467.869 | 5570.270 | 5255.514 | 5148.194 | 7036.126 | 5481.347 | 5567.192 | 5121.827 |
| 24 | V | 428.195 | 287.618 | 572.934 | 1006.373 | 349.682 | 500.024 | 577.988 | 386.138 | 385.297 | 284.353 | 619.899 | 368.103 |
| 25 | OB | 392.990 | 1241.962 | 2327.788 | 315.521 | 361.600 | 781.872 | 560.747 | 901.480 | 264.595 | 571.324 | 804.959 | 512.054 |

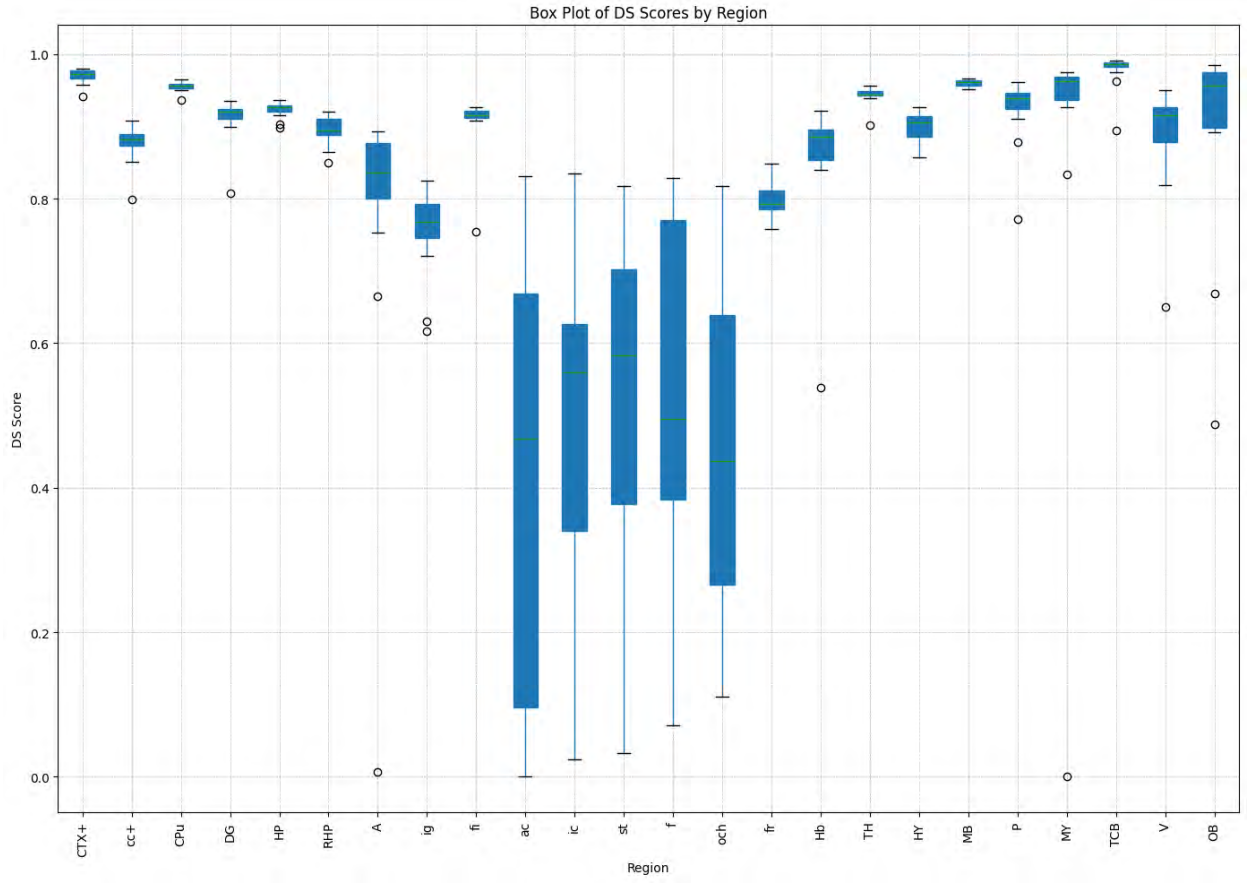



Figure 14: Box Plot of Dice Similarity (DS) Scores by Region Table 1. The box plot illustrates the distribution of DS scores for different brain regions, indicating the model's segmentation accuracy.

cating differences in volume calculation accuracy. Observations include regions like Cortex (CTX+), Corpus Callosum (cc+), Caudate Putamen (CPu), Dentate Gyrus (DG), Hippocampus (HP), and Retro Hippocampus (RHP) showing RAVD values close to 2, with small interquartile ranges (IQRs), suggesting high consistency. In contrast, regions such as Amygdala (A), Indusium Griseum (ig), Fimbria (fi), Anterior Commissure (ac), Internal Capsule (ic), Stria Terminalis (st), and Fornix (f) show more spread-out RAVD values with

Table 3: Region Segmentation Results. Volumetric samples are denoted with the NG prefix, by unique identifiers

| S.No | Region | NG4108 | NG4111 | NG4116 | NG4115 | NG4119 | NG4120 | NG4114 | NG4117 | NG4109 | NG4112 | NG4110 | NG4113 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | background | 0.998 | 0.996 | 0.999 | 0.999 | 0.999 | 0.999 | 0.997 | 0.999 | 0.999 | 0.994 | 0.999 | 0.999 |
| 2 | CTX+ | 0.975 | 0.966 | 0.941 | 0.978 | 0.957 | 0.964 | 0.970 | 0.977 | 0.975 | 0.970 | 0.979 | 0.979 |
| 3 | cc+ | 0.908 | 0.884 | 0.877 | 0.875 | 0.799 | 0.851 | 0.882 | 0.882 | 0.891 | 0.868 | 0.894 | 0.888 |
| 4 | CPu | 0.965 | 0.953 | 0.951 | 0.958 | 0.937 | 0.950 | 0.957 | 0.957 | 0.958 | 0.957 | 0.953 | 0.959 |
| 5 | DG | 0.918 | 0.935 | 0.929 | 0.919 | 0.900 | 0.807 | 0.921 | 0.913 | 0.923 | 0.921 | 0.900 | 0.928 |
| 6 | HP | 0.933 | 0.898 | 0.927 | 0.928 | 0.904 | 0.916 | 0.932 | 0.936 | 0.927 | 0.927 | 0.922 | 0.925 |
| 7 | RHP | 0.850 | 0.876 | 0.894 | 0.921 | 0.893 | 0.865 | 0.914 | 0.910 | 0.895 | 0.902 | 0.911 | 0.892 |
| 8 | A | 0.816 | 0.665 | 0.893 | 0.829 | 0.877 | 0.754 | 0.006 | 0.887 | 0.831 | 0.846 | 0.877 | 0.842 |
| 9 | ig | 0.787 | 0.720 | 0.762 | 0.769 | 0.630 | 0.767 | 0.755 | 0.768 | 0.825 | 0.811 | 0.617 | 0.810 |
| 10 | fi | 0.923 | 0.924 | 0.926 | 0.911 | 0.754 | 0.917 | 0.919 | 0.915 | 0.922 | 0.914 | 0.908 | 0.912 |
| 11 | ac | 0.460 | 0.751 | 0.012 | 0.477 | 0.124 | 0.459 | 0.543 | 0.831 | 0.013 | 0.649 | 0.729 | 0.000 |
| 12 | ic | 0.583 | 0.728 | 0.024 | 0.580 | 0.377 | 0.592 | 0.538 | 0.835 | 0.232 | 0.379 | 0.759 | 0.129 |
| 13 | st | 0.402 | 0.817 | 0.033 | 0.305 | 0.699 | 0.601 | 0.776 | 0.689 | 0.142 | 0.713 | 0.477 | 0.565 |
| 14 | f | 0.310 | 0.828 | 0.407 | 0.372 | 0.770 | 0.387 | 0.495 | 0.770 | 0.071 | 0.778 | 0.549 | 0.496 |
| 15 | och | 0.607 | 0.734 | 0.111 | 0.428 | 0.561 | 0.751 | 0.381 | 0.818 | 0.273 | 0.445 | 0.243 | 0.214 |
| 16 | fr | 0.802 | 0.804 | 0.758 | 0.795 | 0.848 | 0.780 | 0.791 | 0.789 | 0.836 | 0.787 | 0.764 | 0.833 |
| 17 | Hb | 0.887 | 0.898 | 0.844 | 0.896 | 0.538 | 0.840 | 0.856 | 0.894 | 0.900 | 0.922 | 0.875 | 0.883 |
| 18 | TH | 0.943 | 0.939 | 0.956 | 0.948 | 0.902 | 0.947 | 0.946 | 0.942 | 0.951 | 0.942 | 0.942 | 0.950 |
| 19 | HY | 0.911 | 0.881 | 0.905 | 0.926 | 0.857 | 0.874 | 0.921 | 0.905 | 0.909 | 0.887 | 0.921 | 0.906 |
| 20 | MB | 0.960 | 0.952 | 0.961 | 0.962 | 0.954 | 0.964 | 0.963 | 0.961 | 0.967 | 0.954 | 0.957 | 0.966 |
| 21 | P | 0.962 | 0.938 | 0.947 | 0.952 | 0.910 | 0.931 | 0.878 | 0.941 | 0.771 | 0.929 | 0.945 | 0.946 |
| 22 | MY | 0.975 | 0.970 | 0.967 | 0.969 | 0.834 | 0.940 | 0.927 | 0.959 | 0.001 | 0.955 | 0.971 | 0.969 |
| 23 | TCB | 0.989 | 0.986 | 0.990 | 0.986 | 0.962 | 0.988 | 0.985 | 0.989 | 0.895 | 0.975 | 0.991 | 0.987 |
| 24 | V | 0.913 | 0.904 | 0.937 | 0.917 | 0.650 | 0.950 | 0.819 | 0.923 | 0.924 | 0.871 | 0.949 | 0.880 |
| 25 | OB | 0.900 | 0.982 | 0.669 | 0.947 | 0.892 | 0.973 | 0.968 | 0.985 | 0.487 | 0.946 | 0.985 | 0.966 |

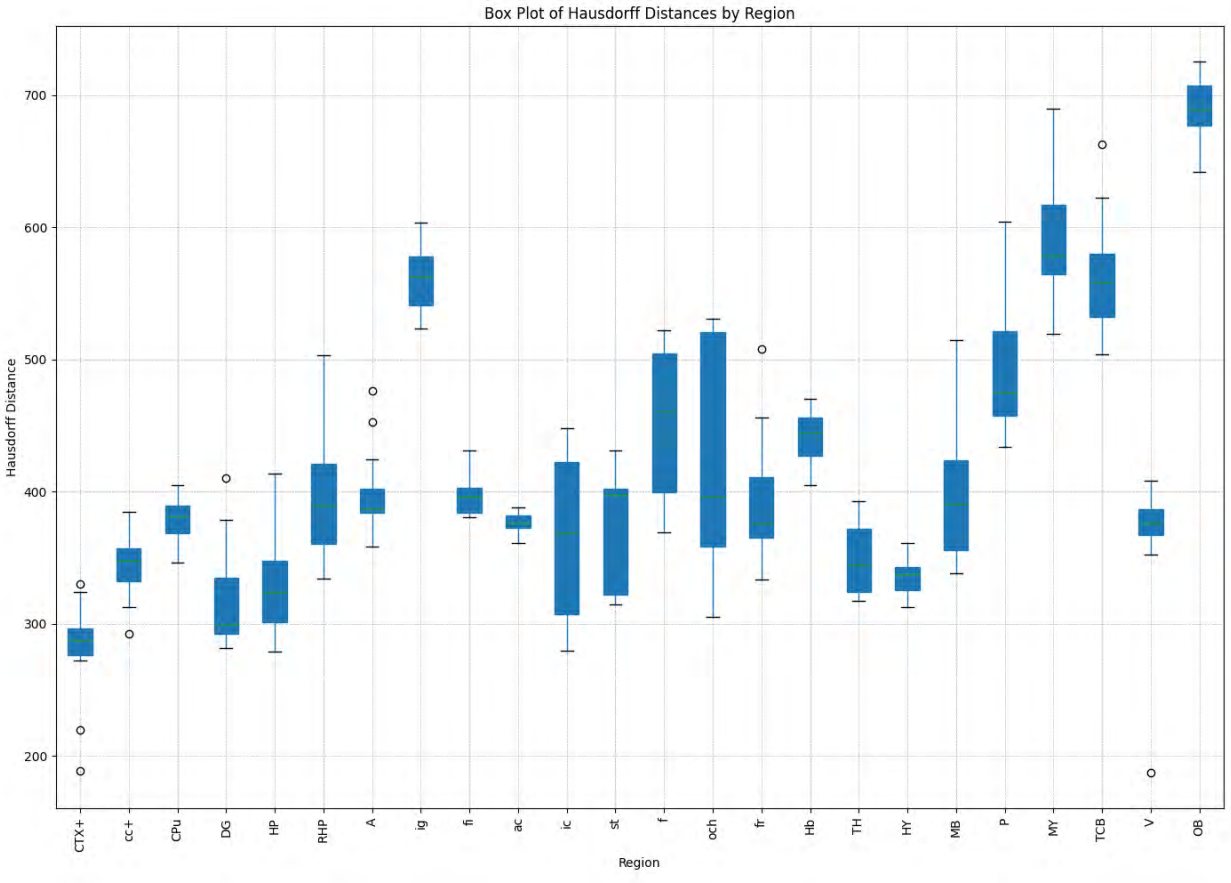



Figure 15: Box Plot of Hausdorff Distance (HD) Scores by Region Table 1.

larger IQRs, indicating more variability in volume differences. Additionally, outliers are observed in several regions, notably in regions like Fimbria (fi), Internal Capsule (ic), and Fornix (f), suggesting occasional significant deviations from the median.

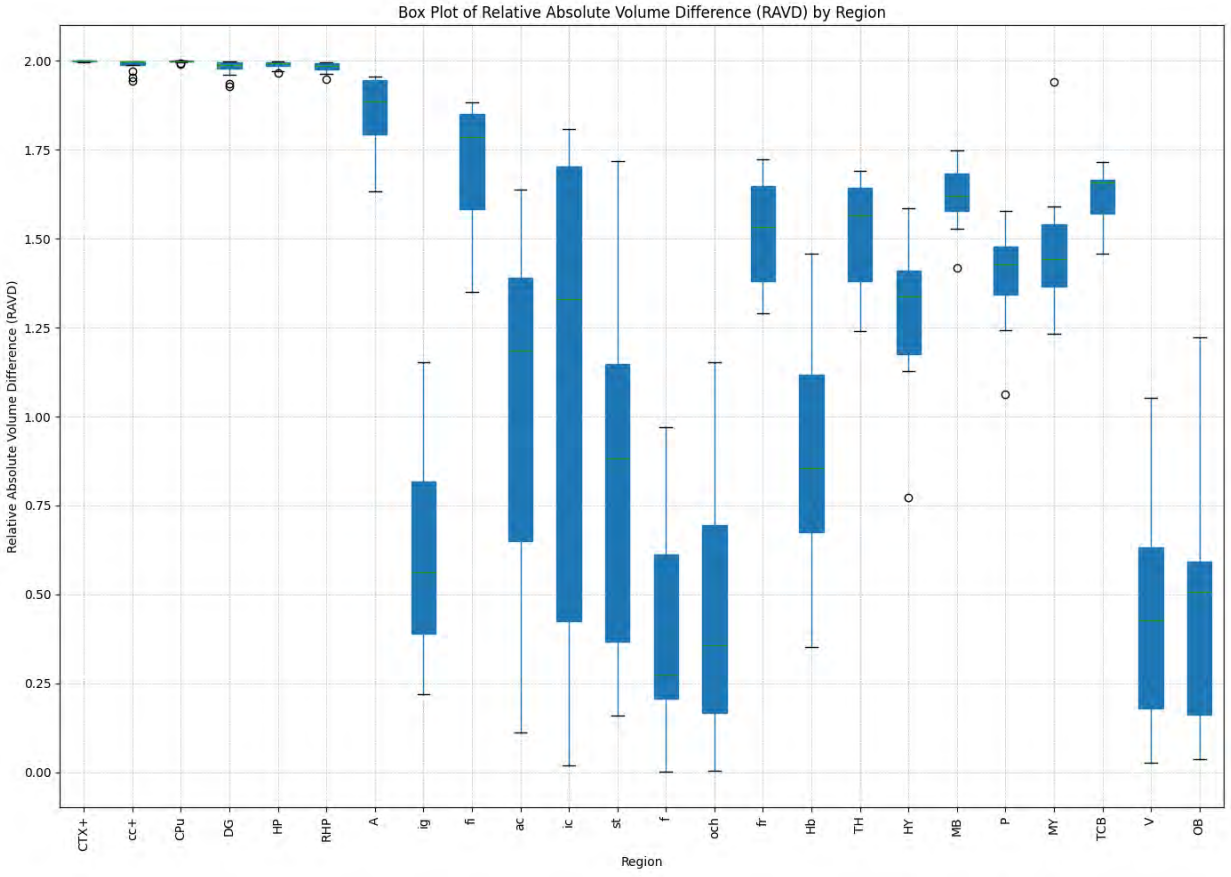



Figure 16: Box Plot of Relative Absolute Volume Difference (RAVD) by Region Table 1.

| | | NG4111 | NG4108 | NG4116 | NG4115 | NG4114 | NG4117 | NG4112 | NG4109 | NG4110 | NG4113 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | backgroun | 0.9958 | 0.9979 | 0.9994 | 0.9992 | 0.9971 | 0.9989 | 0.9935 | 0.9997 | 0.9994 | 0.9993 |
| 2 | CTX+ | 0.9663 | 0.9752 | 0.9409 | 0.9784 | 0.9703 | 0.9771 | 0.9696 | 0.9748 | 0.9786 | 0.9793 |
| 3 | cc+ | 0.8842 | 0.9081 | 0.8773 | 0.8751 | 0.8816 | 0.8824 | 0.8684 | 0.8913 | 0.8938 | 0.8883 |
| 4 | CPu | 0.9531 | 0.9649 | 0.9505 | 0.9583 | 0.9572 | 0.9566 | 0.9572 | 0.9584 | 0.9528 | 0.9592 |
| 5 | DG | 0.9353 | 0.9175 | 0.929 | 0.9194 | 0.9207 | 0.9133 | 0.9211 | 0.9234 | 0.8996 | 0.9279 |
| 6 | HP | 0.8982 | 0.9329 | 0.9271 | 0.928 | 0.9318 | 0.9359 | 0.9267 | 0.9268 | 0.9215 | 0.9248 |
| 7 | RHP | 0.876 | 0.8496 | 0.894 | 0.9205 | 0.9137 | 0.9097 | 0.9018 | 0.8952 | 0.9111 | 0.8922 |
| 8 | A | 0.6654 | 0.8158 | 0.8929 | 0.8289 | 0.0062 | 0.887 | 0.846 | 0.8312 | 0.8771 | 0.8417 |
| 9 | ig | 0.7204 | 0.787 | 0.7617 | 0.7691 | 0.7546 | 0.7676 | 0.8106 | 0.8247 | 0.617 | 0.81 |
| 10 | fi | 0.9242 | 0.9229 | 0.9264 | 0.9113 | 0.9193 | 0.9145 | 0.9138 | 0.9215 | 0.9079 | 0.9123 |
| 11 | f | 0.7509 | 0.4595 | 0.0123 | 0.4768 | 0.5426 | 0.8308 | 0.6486 | 0.0132 | 0.7293 | 0 |
| 12 | st | 0.7283 | 0.5831 | 0.024 | 0.5804 | 0.5382 | 0.8351 | 0.3787 | 0.2315 | 0.759 | 0.1288 |
| 13 | ic | 0.8173 | 0.4018 | 0.0327 | 0.3046 | 0.7759 | 0.6892 | 0.7134 | 0.1423 | 0.4766 | 0.5645 |
| 14 | och | 0.8284 | 0.3095 | 0.4066 | 0.372 | 0.495 | 0.7697 | 0.7777 | 0.0708 | 0.549 | 0.4963 |
| 15 | ac | 0.7343 | 0.6074 | 0.1105 | 0.428 | 0.3812 | 0.8178 | 0.4452 | 0.273 | 0.2432 | 0.2135 |
| 16 | fr | 0.8043 | 0.802 | 0.7575 | 0.7951 | 0.791 | 0.7891 | 0.7869 | 0.8356 | 0.7642 | 0.8328 |
| 17 | Hb | 0.8975 | 0.8873 | 0.8437 | 0.8956 | 0.8561 | 0.8942 | 0.922 | 0.8996 | 0.8745 | 0.883 |
| 18 | TH | 0.9391 | 0.9426 | 0.9559 | 0.948 | 0.9459 | 0.9424 | 0.9424 | 0.9509 | 0.9423 | 0.9495 |
| 19 | HY | 0.8814 | 0.9112 | 0.9052 | 0.9261 | 0.9209 | 0.9054 | 0.8866 | 0.909 | 0.9208 | 0.9061 |
| 20 | MB | 0.952 | 0.96 | 0.9606 | 0.9623 | 0.9631 | 0.9611 | 0.9539 | 0.9668 | 0.9569 | 0.9658 |
| 21 | P | 0.938 | 0.9619 | 0.9469 | 0.9521 | 0.8784 | 0.9407 | 0.9289 | 0.7713 | 0.9452 | 0.9464 |
| 22 | MY | 0.9704 | 0.9753 | 0.9668 | 0.9688 | 0.9265 | 0.9592 | 0.955 | 0.0009 | 0.9705 | 0.9686 |
| 23 | TCB | 0.9859 | 0.9889 | 0.99 | 0.9859 | 0.9854 | 0.9891 | 0.9745 | 0.8946 | 0.9907 | 0.9866 |
| 24 | V | 0.9037 | 0.9134 | 0.9368 | 0.9173 | 0.8187 | 0.923 | 0.8707 | 0.9238 | 0.9494 | 0.8802 |
| 25 | OB | 0.9819 | 0.9001 | 0.6689 | 0.9469 | 0.9681 | 0.9851 | 0.9461 | 0.4872 | 0.985 | 0.9662 |

Equal or less than 0.20
Between 0.20 and 0.40
Between 0.40 and 0.60
Between 0.60 and 0.80
Equal or more than 0.8

Figure 17: 3D full resolution segmentation result of 24 region values of different volume IDs across various samples. The color scale represents the values, where red indicates values equal to or less than 0.20, orange indicates values between 0.20 and 0.40, yellow indicates values between 0.40 and 0.60, light green indicates values between 0.60 and 0.80, and green indicates values equal to or more than 0.80.

| SL | Volume ID | NG4111 | NG4108 | NG4116 | NG4115 | NG4114 | NG4117 | NG4112 | NG4109 | NG4110 | NG4113 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | background | 0.9944 | 0.9973 | 0.9981 | 0.9123 | 0.9949 | 0.9975 | 0.9942 | 0.998 | 0.9984 | 0.9981 |
| 2 | CTX+ | 0.9603 | 0.9724 | 0.9605 | 0.8405 | 0.9614 | 0.973 | 0.9684 | 0.9711 | 0.9761 | 0.9746 |
| 3 | cc+ | 0.8676 | 0.899 | 0.8655 | 0.3942 | 0.8554 | 0.8727 | 0.8496 | 0.8716 | 0.8819 | 0.871 |
| 4 | CPu | 0.948 | 0.9657 | 0.9497 | 0.8063 | 0.9508 | 0.9471 | 0.9522 | 0.9584 | 0.9467 | 0.9556 |
| 5 | DG | 0.93 | 0.8965 | 0.912 | 0.5763 | 0.8967 | 0.9076 | 0.9222 | 0.9127 | 0.8819 | 0.9212 |
| 6 | HP | 0.8999 | 0.9284 | 0.919 | 0.7332 | 0.8992 | 0.9345 | 0.9194 | 0.9286 | 0.9169 | 0.9166 |
| 7 | RHP | 0.8832 | 0.831 | 0.8494 | 0.7415 | 0.8866 | 0.9161 | 0.8882 | 0.8929 | 0.9072 | 0.8793 |
| 8 | A | 0.6847 | 0.8196 | 0.8399 | 0.5735 | 0.0035 | 0.8472 | 0.8068 | 0.8405 | 0.8647 | 0.8407 |
| 9 | ig | 0.5727 | 0.7202 | 0.7531 | 0 | 0.7404 | 0.7555 | 0.76 | 0.7747 | 0.6692 | 0.7987 |
| 10 | fi | 0.9103 | 0.9131 | 0.9185 | 0.5399 | 0.9104 | 0.8974 | 0.8902 | 0.9148 | 0.9078 | 0.9037 |
| 11 | f | 0.7187 | 0.6957 | 0 | 0.0357 | 0.7713 | 0.8949 | 0.7424 | 0.0008 | 0.7744 | 0 |
| 12 | st | 0.7564 | 0.5867 | 0.0149 | 0.4423 | 0.78 | 0.8328 | 0.6299 | 0.0279 | 0.7865 | 0.0303 |
| 13 | ic | 0.8244 | 0.4422 | 0.0095 | 0.1816 | 0.8104 | 0.7168 | 0.7963 | 0.0146 | 0.5953 | 0.036 |
| 14 | och | 0.8894 | 0.6644 | 0 | 0.095 | 0.8525 | 0.9025 | 0.8726 | 0 | 0.6819 | 0 |
| 15 | ac | 0.8329 | 0.729 | 0.009 | 0.0103 | 0.8797 | 0.7922 | 0.6467 | 0.0052 | 0.396 | 0.0037 |
| 16 | fr | 0.7702 | 0.7862 | 0.7568 | 0 | 0.7493 | 0.6653 | 0.7598 | 0.8378 | 0.7247 | 0.8239 |
| 17 | Hb | 0.8752 | 0.8782 | 0.8411 | 0.0969 | 0.828 | 0.8835 | 0.9069 | 0.8914 | 0.8425 | 0.8868 |
| 18 | TH | 0.9392 | 0.9431 | 0.9494 | 0.728 | 0.9386 | 0.9387 | 0.9399 | 0.9493 | 0.9376 | 0.9469 |
| 19 | HY | 0.8773 | 0.9096 | 0.9178 | 0.5383 | 0.917 | 0.9018 | 0.8836 | 0.9096 | 0.9076 | 0.9111 |
| 20 | MB | 0.9533 | 0.957 | 0.9577 | 0.6947 | 0.9594 | 0.9566 | 0.9519 | 0.963 | 0.9544 | 0.9638 |
| 21 | P | 0.9427 | 0.9618 | 0.9446 | 0.7168 | 0.8759 | 0.9469 | 0.9268 | 0.7711 | 0.933 | 0.9403 |
| 22 | MY | 0.9721 | 0.9632 | 0.9183 | 0.7726 | 0.9151 | 0.9545 | 0.9543 | 0.0029 | 0.9612 | 0.9615 |
| 23 | TCB | 0.9848 | 0.9826 | 0.9837 | 0.8398 | 0.9828 | 0.9825 | 0.9764 | 0.8925 | 0.988 | 0.9867 |
| 24 | V | 0.8937 | 0.893 | 0.9326 | 0.2373 | 0.8181 | 0.9054 | 0.8555 | 0.9091 | 0.9316 | 0.87 |
| 25 | OB | 0.9616 | 0.8612 | 0.8102 | 0.3356 | 0.9724 | 0.9796 | 0.9476 | 0.4751 | 0.9805 | 0.9506 |

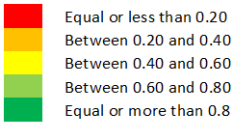



Figure 18: 3D low-resolution segmentation result of 24 region values of different volume IDs across various samples. The color scale represents the values, where red indicates values equal to or less than 0.20, orange indicates values between 0.20 and 0.40, yellow indicates values between 0.40 and 0.60, light green indicates values between 0.60 and 0.80, and green indicates values equal to or more than 0.80.

## 5. Discussion

In our study, the capabilities of several advanced segmentation models to segment 3D volumes of histological mouse brain imaging. The widely recognized U-Net and its derivatives, as well as state-of-the-art frameworks such as nnU-Net, are included. The importance of selecting appropriate preprocessing techniques tailored to specific artifact types is underscored, and future work may see the integration of advanced machine-learning techniques to further enhance artifact removal and image quality. Comprehensive and accurate segmentation across different levels of image detail is ensured by this multi-resolution strategy, making the nnU-Net configuration highly effective for histological mouse brain image analysis research. The experimentation focuses on comparing these models across different resolutions and modalities. Exceptional performance in both 2D and 3D contexts has been demonstrated by nnU-Net. Compelling results were achieved with nnU-Net in three distinct setups: 2D segmentation, 3D segmentation at low resolution, and 3D segmentation at full resolution. The robust adaptability and efficiency of nnU-Net in handling diverse data scales and complexities were highlighted by each of these configurations. The importance of selecting appropriate model configurations and conducting multi-fold validations to ensure robust performance is underscored. Future research may see advanced techniques explored to mitigate variability and further enhance segmentation accuracy and efficiency. The effectiveness of nnU-Net in histological mouse brain analysis is validated by these outcomes, suggesting that it could serve as a benchmark for future developments in the field. As these models continue to be refined, the goal is to further enhance their accuracy and reduce computational demands, potentially leading to faster and more reliable diagnostic tools.

Working with large volumes of data presents significant challenges, particularly in terms of computational time and the analysis of small regions within the dataset. The sheer size of the data can overwhelm processing capabilities, leading to extended computing times that can hinder timely insights and decision-making. Additionally, focusing on small regions within such few number of 3d histological image datasets can be especially difficult, as it requires precise and efficient data-handling techniques to isolate and analyze these areas without losing context or accuracy. This complexity often demands advanced computational resources and sophisticated algorithms to manage, process, and interpret the data effectively, ensuring that the detailed insights required from small regions are not compromised by the overarching volume.

In general, regions Cortex, Corpus Callosum, Caudate Putamen, Dentate Gyrus, Hippocampus, and Retro Hippocampus consistently show better performance across all metrics (lower RAVD and Hausdorff distances, and higher DS Scores), suggesting higher accuracy and reliability in these regions. Regions such as Amygdala, Fimbria, Internal Capsule, and Stria Terminalis exhibit greater variability and lower performance across all metrics, indicating areas that may need further refinement in volume calculation, boundary match-

ing, and segmentation processes. The presence of outliers across several regions suggests occasional deviations that could be due to specific cases or anomalies in data. Overall, the box plots provide a comprehensive view of the performance across different regions, highlighting areas of strength and those requiring improvement.

Building on the automatic solution for segmenting 24 regions in 3D microscopic volumes of mouse brains, future research can be focused on refining the segmentation of smaller, intricate regions. Models can be developed to accurately identify sub-nuclei within larger structures using high-resolution imaging and integrating data from modalities like histological images. Advanced machine learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), will be utilized to enhance segmentation accuracy. Detailed brain atlases and automated quality control mechanisms will be created to ensure reliability. Collaborative platforms for data sharing and the development of user-friendly interfaces will be established to facilitate widespread adoption and advance neuroanatomical and clinical research.

## 6. Conclusions

In this work, we proposed an automatic solution to segment 24 regions of interest in 3D volumetric histological images of mouse brains. The proposed model delivers excellent results across all regions and provides a foundation for developing more precise histological image segmentation systems. The model is user-friendly and does not require any additional software or training for laboratory staff. It processes specific types of histological images and automatically converts them into landmarks for the regions of interest in the mouse brain. The system can accurately identify the 24 regions in as little as 5 minutes, primarily due to validation and correction if needed, compared to an average of 30 hours for manual segmentation. The final output consists of .nrrd files with region labels defined for the analyzed regions, which can be utilized by conventional medical imaging software for various neuroanatomical studies.

## Acknowledgments

First and foremost, I express my deepest gratitude to Almighty Allah for granting me the strength and perseverance to complete this master's thesis.

I extend my heartfelt thanks to my professors and mentors for their invaluable support and guidance throughout my journey. Special thanks to Prof. Fabrice MERIAUDEAU, Prof. Alain Lalande, Prof. Binnaz Yalcin, and Prof. Stephan Collins for their dedication to meeting with me regularly and providing insightful feedback has been instrumental in my academic and personal growth.

I am grateful to Prof. Dr. Siddiqur Rahman, Chairman of Dristi Eye Hospitals, and Dr. Muhammad Moniruzzaman, Chairman of Vision Eye Hospital, for their generous guidance and funding support.

A special thank you to my wife, Shahria Tapa, for her unwavering support, patience, and encouragement.

Lastly, I extend my heartfelt gratitude to all my colleagues at ICMUB, NGMM, the Medical Imaging and Applications (MAIA) Erasmus Mundus Master program, and the affiliated research institutions for providing a conducive and supportive research environment.

Thank you all for your unwavering support and encouragement.

## References

3D Slicer, . 3d slicer. https://www.slicer.org/. Accessed: 2024-06-06.

Biomedisa, . Biomedisa: Biomedical image segmentation app. https://biomedisa.info/. Accessed: 2024-06-06.

BrainVoyager, 2023. Manual segmentation tools, in: BrainVoyager User's Guide. URL: https://www.brainvoyager.com/bv/doc/UsersGuide/Segmentation/ManualSegmentationTools.html.

Cisneros, J., Lalande, A., Yalcin, B., Meriaudeau, F., Collins, S., 2023. Automatic segmentation of histological images of mouse brains 16, 553. URL: https://www.mdpi.com/1999-4893/16/12/553, doi:10.3390/a16120553. number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

Collins, S.C., Mikhaleva, A., Vrcelj, K., Vancollie, V.E., Wagner, C., Demeure, N., Whitley, H., Kannan, M., Balz, R., Anthony, L.F.E., Edwards, A., Moine, H., White, J.K., Adams, D.J., Reymond, A., Lelliott, C.J., Webber, C., Yalcin, B., 2019. Large-scale neuroanatomical study uncovers 198 gene associations in mouse brain morphogenesis 10. URL: https://www.nature.com/articles/s41467-019-11431-2, doi:10.1038/s41467-019-11431-2. publisher: Nature Publishing Group.

Isensee, F., Jaeger, P.F., Kohl, S.A.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. doi:10.1038/s41592-020-01008-z.

Jones, C., Liu, T., Cohan, N.W., Ellisman, M., Tasdizen, T., 2015. Efficient semi-automatic 3d segmentation for neuron tracing in electron microscopy images. Journal of Neuroscience Methods 246, 13–21. doi:10.1016/j.jneumeth.2015.03.005.

Jubair, I., R., M., 2023. Loss functions.... dice score in inference. URL: https://github.com/MIC-DKFZ/nnUNet/issues/1792. accessed: 2024-06-05.

Kretz, P.F., Wagner, C., Mikhaleva, A., Montillot, C., Hugel, S., Morella, I., Kannan, M., Fischer, M.C., Milhau, M., Yalcin, I., Brambilla, R., Selloum, M., Herault, Y., Reymond, A., Collins, S.C., Yalcin, B., 2023. Dissecting the autism-associated 16p11.2 locus identifies multiple drivers in neuroanatomical phenotypes and unveils a male-specific role for the major vault protein 24, 261. URL: https://doi.org/10.1186/s13059-023-03092-8, doi:10.1186/s13059-023-03092-8.

Ma, J., . Segment anything in medical images .

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation URL: http://arxiv.org/abs/1505.04597, doi:10.48550/arXiv.1505.04597, arXiv:1505.04597 [cs].

Scheenstra, A.E., van de Ven, R.C., van der Weerd, L., van den Maagdenberg, A.M., Dijkstra, J., Reiber, J.H., 2009. Automated seg-

mentation of in vivo and ex vivo mouse brain magnetic resonance images. Molecular Imaging 8, 35–44.

Tappan, S.J., Eastwood, B.S., O'Connor, N., Wang, Q., Ng, L., Feng, D., Hooks, B.M., Gerfen, C.R., Hof, P.R., Schmitz, C., Glaser, J.R., 2019. Automatic navigation system for the mouse brain. Journal of Comparative Neurology 527, 1454–1465.

Uberti, M.G., Boska, M.D., Liu, Y., 2009. A semi-automatic image segmentation method for extraction of brain volume from in vivo mouse head magnetic resonance imaging using constraint level sets 179, 338–344. URL: https://www.sciencedirect.com/science/article/pii/S0165027009001083, doi:10.1016/j.jneumeth.2009.02.007.

Vezhnevets, V., Konouchine, V., 2005. Growcut: Interactive multi-label nd image segmentation by cellular automata, in: proc. of Graphicon, Citeseer. pp. 150–156.

Zhu, L., Kolesov, I., Gao, Y., Kikinis, R., Tannenbaum, A., 2014. An effective interactive medical image segmentation method using fast growcut, in: MICCAI workshop on interactive medical image computing.

## Appendix

### A. Segment Anything Model (SAM)

**Segment Anything in Medical Images (MedSAM)** In the 2D space, MedSAM performs segmentation tasks by processing 3D medical images such as CT and MRI scans as a series of 2D slices. This approach simplifies the segmentation process and allows for the effective application of 2D segmentation techniques uniformly across various types of images. Bounding box prompts are utilized to specify the region of interest (ROI) that needs to be segmented.

The process involves user interaction where the user draws a bounding box around the area of interest in the 2D slice, providing spatial context and helping the model focus on the specific region. The coordinates of the bounding box are transformed into a feature representation using positional encoding, which is then fed into the model's prompt encoder. The image and the encoded bounding box prompt are passed through the model, where the image encoder extracts high-dimensional features from the image, and the prompt encoder processes the bounding box information. The mask decoder fuses the image features and the prompt features using cross-attention mechanisms to generate the final segmentation mask, highlighting the segmented area within the bounding box.

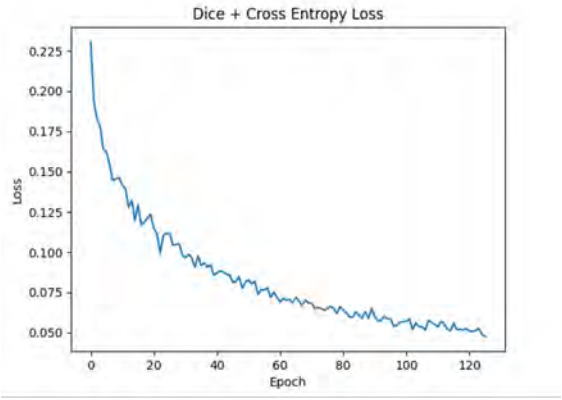



Figure 19: MesSAM training.

Point prompts involve placing specific points within the ROI to guide the segmentation process. The user marks points on the image to indicate areas to include or exclude in the segmentation, using positive points (inside the ROI) and negative points (outside the ROI) to refine the segmentation boundaries. Each point is encoded into a high-dimensional space, capturing its position relative to the image, and these encodings are processed by the prompt encoder. The model receives the image and the encoded point prompts, with the image encoder extracting relevant features and the prompt encoder integrating the point information. The

mask decoder combines the image features with the point prompt features using cross-attention mechanisms to generate a segmentation mask. These points help the model understand the exact regions to include or exclude, allowing for fine-tuning of the segmentation boundaries Ma.

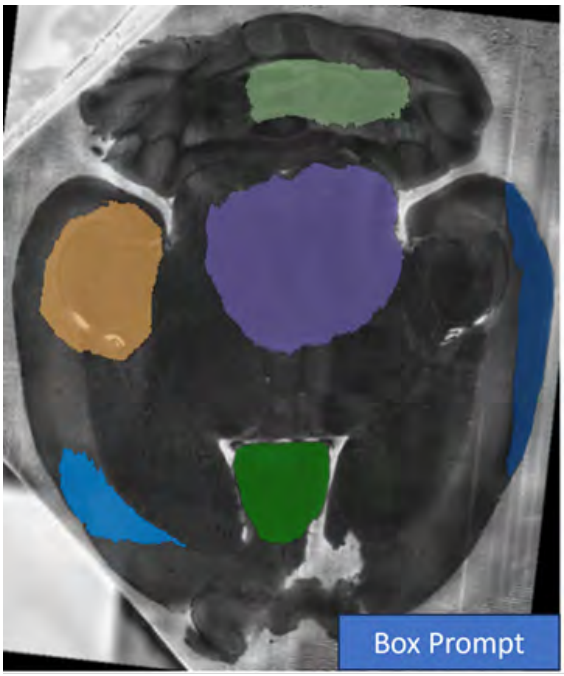



Figure 20: Segmentation Results Using medSAM Box Prompts on 2D Slice Images.

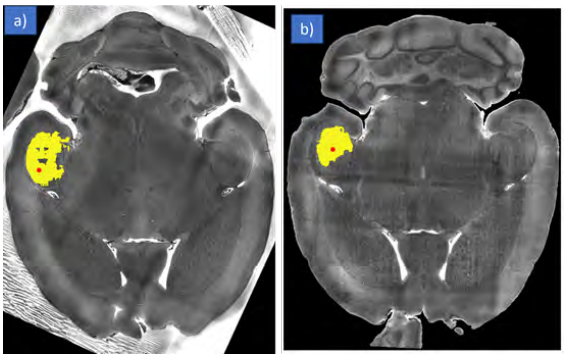



Figure 21: Comparison of medSAM Point Prompts on 2D Slice Images: (a) Segmentation results without preprocessing, (b) Segmentation results with preprocessing. The preprocessing step enhances the accuracy and clarity of the segmentation.

**MaskSAM Framework** The MaskSAM framework introduces several key innovations to adapt SAM from 2D natural images to 3D medical images. The prompt generator is employed in conjunction with SAM's image encoder to generate auxiliary classifier tokens, binary masks, and bounding boxes, thereby eliminating the need for manual prompts. The 3D Depth-Convolution Adapter (DConvAdapter) is designed for image embeddings to enable the extraction of 3D information, while the 3D Depth-MLP Adapter (DMLPAdapter) is tailored for prompt embeddings to manage the additional depth dimension in medical images.

Several modifications were made to SAM's image encoder and mask decoder to cater to the unique requirements of 3D medical images. In the image encoder, a sequence of convolutional layers for channel adaptation was added, allowing varied modalities of medical images to be processed in the RGB channels used by SAM. A learnable depth positional embedding was introduced to better understand the depth information in medical images. Additionally, DConvAdapter blocks were inserted into each attention block of the image encoder to enhance the understanding of 3D spatial relationships.

In the mask decoder, learnable global classifier tokens were added to predict semantic labels for each binary mask, along with a learnable depth positional embedding to capture depth information. DMLPAdapter and DConvAdapter blocks were integrated into appropriate locations within the decoder to effectively process depth information.

Furthermore, a dataset mapping pipeline was developed to convert multi-class masks into sets of binary masks with semantic labels. This conversion process decomposes multi-class masks into binary masks for each class. Bipartite matching is utilized between the predicted masks and ground truth segments to select the best-matching predictions for loss calculation.

In our study, the results obtained from segmenting 2D histological images using the MedSAM model were not satisfactory. We utilized bounding box prompts to provide spatial context and point prompts to mark specific points within regions of interest. Initially, we tested the model without any preprocessing, which did not yield satisfactory results. Subsequently, we applied preprocessing techniques such as normalization and noise reduction to the images before training the model. Despite these efforts, the segmentation results remained unsatisfactory. These challenges highlight the need for further optimization or alternative methods to achieve better accuracy in 2D histological image segmentation. The model struggled to accurately delineate the regions of interest, often resulting in imprecise boundaries and misclassification of anatomical structures. This suggests that while MedSAM shows potential, it may require further optimization and fine-tuning for 2D histological image segmentation tasks. The complexity and variability inherent in histological images pose significant challenges that the current model iteration could not adequately address. Consequently, our focus has shifted towards exploring 3D volumetric approaches and advanced machine learning techniques to achieve better accuracy and reliability in segmentation outcomes.

## B. Segmented full brain 24 regions

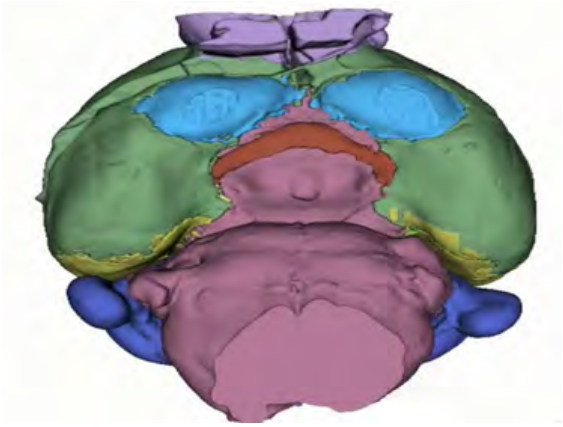


Figure 22: Segmented full brain 24 regions Table 1.

## C. Small region identification of histological mouse brain by pixel counts

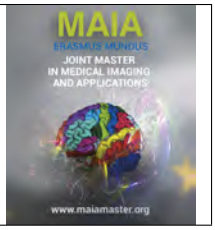
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Region | NG4108 | NG4111 | NG4116 | NG4119 | NG4115 | NG4120 | NG4114 | NG4117 | NG4109 | NG4112 | NG4110 | NG4113 |
| 2 | BG | 100013582 | 52176648 | 94258667 | 59210298 | 70221045 | 77420685 | 52349698 | 55401809 | 61268942 | 47907029 | 98251432 | 82188426 |
| 3 | CTX+ | 13123790 | 12031753 | 13780858 | 11017591 | 15015588 | 10144349 | 9309837 | 11340841 | 15066101 | 13047163 | 14989756 | 13846545 |
| 4 | cc+ | 908045 | 988205 | 999139 | 728859 | 1061451 | 809458 | 737164 | 901475 | 1030712 | 1103185 | 1130853 | 1071331 |
| 5 | CPu | 5677092 | 5888111 | 5242927 | 4939545 | 5739518 | 5731060 | 5049717 | 5393050 | 5701843 | 5799007 | 5882564 | 5793639 |
| 6 | DG | 686849 | 721753 | 597888 | 536471 | 671236 | 795151 | 597663 | 590658 | 671340 | 700306 | 652725 | 713284 |
| 7 | HP | 1506707 | 1580086 | 1460669 | 1310951 | 1706985 | 1362181 | 1423286 | 1425295 | 1611106 | 1734336 | 1810564 | 1818336 |
| 8 | RHP | 1297884 | 1208729 | 1594038 | 1240692 | 1722468 | 1143885 | 1148011 | 1382542 | 1603376 | 1648532 | 1654333 | 1435451 |
| 9 | A | 250754 | 182951 | 284201 | 189176 | 300290 | 172005 | 664 | 224049 | 238222 | 274419 | 274820 | 282921 |
| 10 | ig | 10214 | 9567 | 11761 | 13625 | 13165 | 7224 | 5478 | 10612 | 10908 | 8309 | 13005 | 10483 |
| 11 | fi | 272023 | 289190 | 268655 | 168731 | 279582 | 234104 | 250916 | 240480 | 282728 | 288411 | 281690 | 304269 |
| 12 | ac | 109093 | 93982 | 71913 | 65628 | 77476 | 149950 | 71942 | 118400 | 110986 | 72515 | 136343 | 80568 |
| 13 | ic | 365518 | 80659 | 80808 | 48385 | 407570 | 470984 | 67236 | 361389 | 414435 | 60174 | 406359 | 72852 |
| 14 | st | 77500 | 397402 | 411726 | 341105 | 66444 | 53237 | 313471 | 65055 | 66817 | 429097 | 72708 | 336290 |
| 15 | f | 66535 | 146450 | 137898 | 131135 | 65727 | 37453 | 104768 | 49852 | 70986 | 100821 | 65740 | 138812 |
| 16 | och | 80672 | 79258 | 60538 | 44736 | 95518 | 77056 | 50204 | 91356 | 99020 | 101268 | 86433 | 70648 |
| 17 | fr | 19389 | 18241 | 18111 | 19954 | 19267 | 19807 | 22124 | 15584 | 22303 | 23305 | 16784 | 25441 |
| 18 | Hb | 86286 | 80434 | 83132 | 37666 | 76835 | 89292 | 71064 | 72663 | 82581 | 78683 | 69167 | 76541 |
| 19 | TH | 2087008 | 2115503 | 1892110 | 1447076 | 1945981 | 1766101 | 1620322 | 1728938 | 1957220 | 2002845 | 2098648 | 2160638 |
| 20 | HY | 1445082 | 1474367 | 1543971 | 1295246 | 1422002 | 1784889 | 1385987 | 1411380 | 1567157 | 1253825 | 1632031 | 1431069 |
| 21 | MB | 3766019 | 3812262 | 3318945 | 3457845 | 3460886 | 3790082 | 3478057 | 3424259 | 3518657 | 3753363 | 3649733 | 3762247 |
| 22 | P | 2427230 | 2336754 | 2262427 | 2324468 | 2271670 | 2440851 | 2597559 | 2314373 | 3791706 | 2343732 | 2528145 | 2522491 |
| 23 | MY | 3055420 | 3278200 | 2907002 | 2540666 | 3225173 | 3193210 | 2903537 | 2630564 | 7674 | 3417776 | 2991147 | 3246510 |
| 24 | TCB | 5527757 | 5066770 | 4999864 | 5434437 | 5467869 | 5570270 | 5255514 | 5148194 | 7036126 | 5481347 | 5567192 | 5121827 |
| 25 | V | 428195 | 287618 | 572934 | 1006373 | 349682 | 500024 | 577988 | 386138 | 385297 | 284353 | 619899 | 368103 |
| 26 | OB | 392990 | 1241962 | 2327788 | 315521 | 361600 | 781872 | 560747 | 901480 | 264595 | 571324 | 804959 | 512054 |

**1. Micro**: 0.66k to 70.99k
**2. Small**: 70.99k to 379.68k
**3. Medium**: 379.68k to 1,298.83k
**4. Big**: 1,298.83k to 2,440.85k
**5. Large**: 2,440.85k to 98,251.43k

Figure 23: Small region identification of histological mouse brain by pixel counts

Medical Imaging and Applications

Master Thesis, June 2024




# Interactive Deep Learning-Based Active Learning Strategies for Abdominal Organ Segmentation

Taofik Ahmed Suleiman[a,b], Joseph Y. Lo[a]

[a] *Center for Virtual Imaging Trials, Department of Radiology, Duke University School of Medicine, Durham, NC, USA*
[b] *Department of Medical Imaging and Computing, University of Girona, Girona, Spain*

**Abstract**

Active learning strategies have emerged as a powerful tool to improve the efficiency and accuracy of segmentation models in medical imaging, particularly when dealing with limited labeled data. In this study, we explored the effectiveness of several active learning methods, including variance-based, entropy-based, and learning loss strategies, in enhancing the performance of a UNETR-based segmentation model for female abdominal organ segmentation. The dataset comprised CT scans with 24 annotated abdominal organs, including the uterus, which was segmented using nnU-Net. We then implemented an interactive active learning framework that integrates multiple strategies and deep learning models for segmentation and automatic/manual label correction. Our findings indicate that active learning strategies, especially the learning loss strategy, significantly outperformed random selection. This was demonstrated by higher Dice scores with fewer iterations. For example, by the second iteration, the learning loss strategy achieved a Dice score of 0.67, compared to 0.60 for random selection. This early advantage highlights the efficiency of active learning in quickly improving model performance. Overall, the learning loss strategy maintained superior performance throughout the iterations, reaching a Dice score of 0.78 by the seventh iteration. The variance and entropy-based methods also showed notable improvements over random selection, though they tended to plateau as the model gained confidence. These results emphasize the importance of using advanced active learning techniques to enhance model performance while reducing annotation costs in abdominal organ segmentation.

*Keywords:* Monai Label, active learning, variance, entropy, learning loss, deep learning, organ segmentation, computed tomography

## 1. Introduction

Segmentation is a technique of partitioning an image into sub-constituent parts, allowing for the extraction of useful information. This process is crucial in image analysis (Mazurowski et al., 2023; Ramesh et al., 2021). In medical imaging, several techniques have been developed for segmentation (Chen et al., 2022; Kim et al., 2020; Liang and Huang, 2018; Luo et al., 2021; Wang et al., 2022; Xiao et al., 2018), particularly organ segmentation which is essential for numerous applications, including computer-aided diagnosis, computer-aided surgery, and radiation therapy (Shimizu et al., 2010; Wolz et al., 2013). Segmentation of internal structures, such as abdominal organs (liver, spleen, colon, uterus, etc.), presents particular challenges due to the abdomen's many soft tissue organs, which often have low contrast from each other, heterogeneous shapes, and complexities arising from dynamic range changes due to air or exogenous contrast material. Thus, accurate segmentation of these organs is critical for various clinical and research applications, including diagnosis, treatment planning, and surgical navigation. However, the manual annotation of these structures is labor-intensive, time-consuming, and subject to inter-observer variability, and as a result, a need for automatic or semi-automatic segmentation methods and that's where deep learning (DL) comes into play (Diaz-Pinto et al., 2024; Hesamian et al., 2019; Razzak et al., 2018).

For automated medical image segmentation, deep learning with convolutional neural networks (CNNs) has achieved state-of-the-art results (Litjens et al., 2017; Wang et al., 2018a,b). Despite intensive studies on deep

learning approaches for automatic or semi-automatic segmentation, there remain challenges that need to be overcome before these methods can be applied to clinical environments. Specifically, the abdominal multi-organ segmentation of computed tomography (CT) images is a significant problem in medical image processing since the distribution and shape of the abdominal organs can vary significantly over time within an individual as well as throughout the population. While continuous integration of novel datasets into the training set provides the potential for better segmentation performance, large-scale data collection is not only expensive but also impractical in some situations because labeled data are valuable resources that can be very expensive to obtain (Rajchl et al., 2016; Wang et al., 2018b). Furthermore, it is uncertain what marginal value additional data have to offer (Xu et al., 2020). To address this problem, several researchers have leveraged the power of active learning techniques, thereby reducing the amount of data to be annotated by radiologists.

In contrast to passive machine learning, active learning (AL) is a special abstraction of machine learning approaches where the model/algorithm could direct users to a group of data points that would be useful to the model if annotated (Nath et al., 2020). In comparison to "passive learning," which is based on random sampling, AL has been demonstrated to perform better in other domains by using fewer annotated examples to get a comparable level of performance (Xu et al., 2020). AL, combined with DL, allows for the development of a framework in which the deep network architecture is coupled with classical techniques to evaluate uncertainty for the selection of samples. Due to its targeted selection of data points that may be described as hard instances, active learning data selection for a model has the potential to accelerate convergence, raise performance with fewer data, and improve robustness (Gal and Ghahramani, 2016; Gal et al., 2017; Nath et al., 2020; Sourati et al., 2018; Yang et al., 2017). Nonetheless, as the diversity of the dataset typically affects AL performance, more research on various AL methodologies is still necessary. To examine several AL methods, we implemented AL strategies based on entropy, variance, and learning losses and compared these techniques with the random selection methods.

## 2. State of the art

Active learning has gained attention in medical image segmentation, especially for addressing the challenge of limited annotated data. Numerous studies have been conducted on AL across various domains. This section presents some of the current research in AL strategies.

### *2.1 Diminishing Uncertainty within the Training Pool*

Nath et al. (Nath et al., 2020) proposed a framework that utilizes a query-by-committee approach for AL, introducing three new strategies: increasing the frequency of uncertain data, using mutual information among input images as a regularizer, and adapting Dice log-likelihood for Stein variational gradient descent (SVGD). These strategies significantly reduce the amount of data needed to achieve full accuracy. The study explores the benefits of active learning specifically for the segmentation of medical imaging datasets, using MRI scans of the hippocampus and CT scans of the pancreas and tumors. Their results indicate an improvement in terms of data reduction, achieving full accuracy while using only 22.69% and 48.85% of the available data for each dataset, respectively (Nath et al., 2020; Tharwat and Schenck, 2023). However, the key limitation of this study is computational cost as the total training time for a single active learning method with 40 active iterations is approximately 160 and 60 GPU hours for pancreas and hippocampus datasets, respectively.

### *2.2 Deep Bayesian Active Learning, and Monte Carlo Dropout*

Deep Bayesian Active Learning, leverages Bayesian neural networks (BNNs) to model uncertainty in predictions. This approach prioritizes samples with the highest uncertainty for labeling, effectively reducing the required labeled dataset size while maintaining high performance. Utilizing dropout as a Bayesian approximation, the method employs Monte Carlo (MC) dropout to estimate model uncertainty by performing multiple stochastic forward passes through the network and calculating the variance in predictions. This technique allows for the creation of an ensemble of neural network models, approximating the posterior distribution of model parameters without significant computational overhead (Gal and Ghahramani, 2016; Gal et al., 2017).

They demonstrated the efficacy of their approach on various tasks, including regression and classification. For instance, applying their method to the MNIST dataset for digit classification showed significant improvements in predictive performance. The use of MC dropout allowed the model to identify uncertain predictions, which could then be prioritized for human labeling, enhancing accuracy with fewer labeled samples. In regression tasks, such as predicting atmospheric CO2 concentrations, the Bayesian approach provided more reliable uncertainty estimates, leading to better generalization on unseen data (Gal et al., 2017).

Other methods have used Monte Carlo dropout as well, Górriz et al., used Monte Carlo dropout to model uncertainty in the melanoma segmentation task (Gorriz et al., 2017), Saidu and Csató applied Monte Carlo dropout in Bayesian UNet for semantic image segmentation (Saidu and Csató, 2021), Xie et al., developed Entropy-Guided Contrastive Learning (EGCL-Net), which combines Monte Carlo dropout with entropy-based methods to improve semi-supervised im-

age segmentation performance (Xie et al., 2024). The major limitation of the Bayesian active learning approach revolves around the potential to fall into local optima if the system is not reset and this results in higher training time. Also, Monte Carlo simulation is generally very expensive in terms of computation time.

### *2.3 Active Learning With Entropy*

Siddiqui et al., proposed ViewAL, a novel active learning strategy for semantic segmentation that exploits viewpoint consistency in multi-view datasets. The core idea is that inconsistencies in model predictions across different viewpoints provide a reliable measure of uncertainty. This method introduces a new viewpoint entropy formulation to quantify these inconsistencies. Additionally, it incorporates superpixel-level uncertainty computations, leveraging localized signals in the segmentation task to reduce annotation costs. By focusing on areas where model predictions vary significantly across viewpoints, ViewAL effectively identifies the most uncertain and informative samples for labeling. The experimental results demonstrated that ViewAL significantly reduces the amount of labeled data required to achieve high performance. For instance, ViewAL achieved 95% of the maximum achievable performance using only 7% of labeled data on the SceneNet-RGBD dataset, compared to the 14% required by the best state-of-the-art method (Siddiqui et al., 2020). This method is generally dependent on multi-view datasets and availability of such data is a key limitation.

Other methods that utilize entropy-based strategies in active learning include Minimax Active Learning by Ebrahimi et al., which combines uncertainty and diversity in an adversarial manner to select samples with high entropy for labeling (Ebrahimi et al., 2020), and the entropy-based active learning approach for object detection by Wu et al., which balances computational complexity with informative sample selection using an Entropy-based Non-Maximum Suppression (ENMS) strategy (Wu et al., 2022).

### *2.4 Active Learning with Stochastic Batches*

Gaillochet et al. developed a stochastic batch querying (SBQ) strategy to enhance uncertainty-based active learning (AL) methods for medical image segmentation. By computing uncertainty at the batch level, SBQ effectively selects diverse and informative samples, improving model performance with reduced annotation effort. Experiments on datasets like PROMISE12 and the Medical Segmentation Decathlon demonstrated that SBQ consistently outperforms traditional uncertainty-based methods, improving metrics such as the Dice similarity coefficient and Hausdorff distance. This approach leverages both random sampling diversity and uncertainty informativeness, reducing redundancy in sample selection but could be very computationally expensive (Gaillochet et al., 2023).

### *2.5 Other Approaches in Active Learning for Medical Image Segmentation*

Zhao et al. proposed DSAL, a deep active semi-supervised learning framework combining active learning and semi-supervised strategies to optimize the use of labeled and unlabeled data (Zhao et al., 2021). Burmeister et al. conducted a comprehensive comparison of various AL strategies for 3D medical image segmentation on the Medical Segmentation Decathlon datasets, providing valuable insights into the strengths and weaknesses of different techniques (Burmeister et al., 2022). Wu et al. developed COWAL, a correlation-aware active learning strategy for surgery video segmentation, which effectively selects representative images from local clusters through a fine-tuned latent space (Wu et al., 2024). Also, Wang et al. in 2019 developed a two-step query method for active learning in medical image segmentation, which calculates sample complexity and potential value to improve segmentation tasks like bladder segmentation (Wang et al., 2019). Yang et al. presented a suggestive annotation framework combining fully convolutional networks with active learning to significantly reduce annotation effort by focusing on the most uncertain and representative areas for annotation (Yang et al., 2017). Additionally, Arikan et al. proposed a deep active learning framework incorporating uncertainty metrics and similarity measures to enhance AL strategies, achieving faster learning and improved robustness in biomedical segmentation tasks (Arikan et al., 2023). These diverse and innovative approaches highlight the potential of active learning to reduce annotation costs and improve model performance in medical image segmentation, however, there is still relatively limited literature on AL work for medical image segmentation compared to classification tasks.

### *2.6 Contributions*

The primary contribution of this work is to implement an interactive active learning framework with uncertainty strategies based on entropy, variance, and learning loss and compare the results to the traditional random selection methods. This framework aims to improve the performance of abdominal organ segmentation while reducing annotation costs. Specifically, it demonstrates that a loss function can be trained and integrated into the training loop of a segmentation task while also serving as an active learning data selection strategy to improve the segmentation model. This approach reduces computation time as compared to other strategies, as loss is easy and fast to compute. The proposed framework addresses the high computational costs associated with other methods by offering a more efficient way to enhance model performance and reduce

the amount of labeled data needed for accurate segmentation, thereby making it feasible for practical applications in medical imaging.

## 3. Material

This section presents the source of the dataset, the preprocessing steps, uterus segmentation, and the methodology for collecting and merging individual labels to form the desired classes for the task.

### *3.1. Dataset*

The dataset comprises 1083 CT whole-body volumes from patients at Duke Hospital with 248 female cases, along with individual masks of over 120 organs. These masks were initially generated using the TotalSegmentator model (Wasserthal et al., 2023) and subsequently post-processed with various techniques, such as those described in (Mouheb et al., 2023) for colon refinement, and verified by a radiologist. For this study, we selected 23 organ masks, focusing on structures within the abdomen. Notably, the uterus was missing from these structures, and as our study focuses on female cases, the inclusion of the uterus is critical. Table 1 lists the abdominal organs included in this study with the label tagged with * not included in the female cases.

Table 1: Full list of the abdominal organ labels used

| ID | Label | ID | Labels |
|---|---|---|---|
| 0 | background | 13 | iliac vena left |
| 1 | seminal vesicles * | 14 | iliac vena right |
| 2 | Rectum * | 15 | inferior vena cava |
| 3 | prostate | 16 | kidney left |
| 4 | adrenal gland left | 17 | kidney right |
| 5 | adrenal gland right | 19 | liver |
| 6 | aorta | 19 | pancreas |
| 7 | colon | 20 | portal vein and splenic vein |
| 8 | duodenum | 21 | small bowel |
| 9 | esophagus | 22 | spleen |
| 10 | gallbladder | 23 | stomach |
| 11 | iliac artery left | 24 | uterus |
| 12 | iliac artery right | | |

### *3.2. Uterus Segmentation*

To incorporate the uterus alongside other available masks, the radiologist annotated about 50 cases, enabling us to train a model for accurate uterus segmentation. Figure 1 shows the pipeline for this process.

#### *3.2.1. Preprocessing*

As a starting point, we preprocessed the CT volumes by removing unnecessary slices along the axial plane. The slices were bounded by the liver and femur, which served as our reference points. This range was chosen because slices outside this region do not contain the uterus and are irrelevant to our task. By excluding these extraneous slices, we significantly reduced the number of irrelevant inputs fed into the model. This step was crucial in ensuring that the training data was more focused, reducing noise and improving the model's ability to distinguish the uterus from other structures. This reduction not only streamlined the dataset for the model to learn to segment the uterus more efficiently and accurately but was also important in dataset optimization.

#### *3.2.2. Model Training with nnU-Net*

The nnU-Net by (Isensee et al., 2019) is an automated deep-learning framework designed for biomedical image segmentation. The nnU-Net standardizes the entire segmentation process, including network architecture, training, preprocessing, and postprocessing pipelines, adapting these components based on the dataset's characteristics. This framework has demonstrated state-of-the-art performance in various medical imaging tasks, particularly in abdominal organ segmentation. For these reasons, we used the nnU-net for segmenting the uterus.

#### *3.2.3. Postprocessing*

Following the segmentation, the images were reconstructed back to their original structure by reversing the initial preprocessing steps. This postprocessing phase was essential to prepare the segmented uterus for integration with other abdominal organ labels and maintain the integrity of the entire abdominal structure.

### *3.3. Label Integration*

At this stage, we have obtained all the desired 24 labels necessary for training the active learning model. However, these labels exist as separate masks, each representing a different organ or structure. To proceed, it is essential to combine these individual masks into a single multi-class segmentation label. To achieve this, We selected one of the masks to serve as the reference. This reference mask provides the spatial properties (such as origin, spacing, and direction) that all other masks must align with. Each additional mask was added to the reference mask one by one. This means that before adding a mask to the reference, we checked for any discrepancies in spatial properties. If a mask did not match the reference in terms of origin, spacing, or direction, it was resampled to align perfectly with the reference mask. This step ensured consistency across all masks, facilitating accurate merging. In addition, each organ mask was assigned a unique label value by multiplying the binary mask with its corresponding label and this created a labeled image for each organ. The result of this process was a single multi-class segmentation image, where each voxel was labeled according to the organ it belonged to. This label integration ensured that all labels were accurately represented as a unified segmentation mask containing 25 classes (including background) necessary for the active learning model stage.

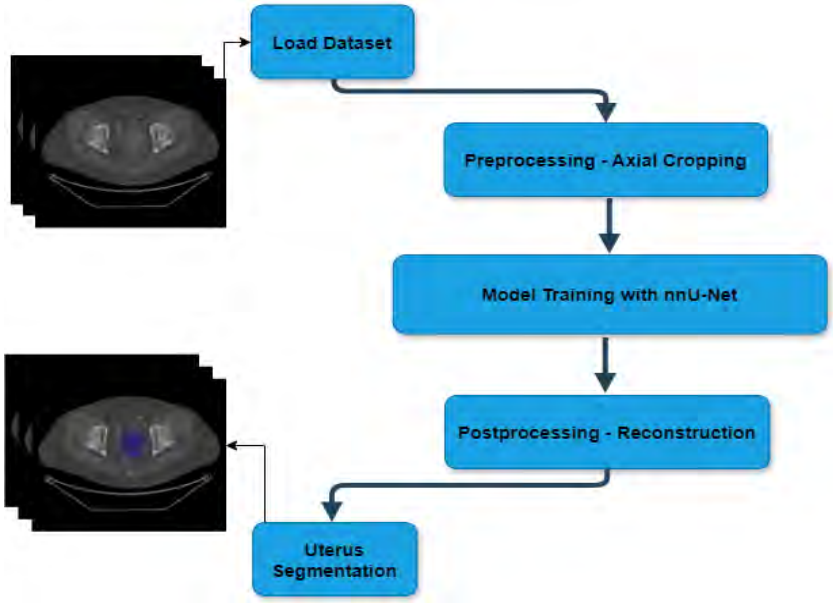



Figure 1: Uterus segmentation workflow

## 4. Method

Our methodology involves integrating MONAI Label with the 3D Slicer to provide an interactive segmentation tool inspired by the authors in (Diaz-Pinto et al., 2024), enabling radiologists to visualize and refine segmentation results in real time. This enables radiologists to review segmentation results and make corrections directly within the 3D Slicer environment during the active learning process. This interactivity ensures that any inaccuracies can be promptly addressed before the next iteration. The Radiologist has the option to manually edit segmentations or use automated correction tools. The DeepEdit and DeepGrow models facilitate this by allowing users to click on points representing the foreground and background of the organ, thereby refining the segmentation. All these features are embedded in the 3D slicer using MONAI Label. Several AL strategies are also implemented on how the data is selected for the active learning phase. Figure 2 illustrates the proposed pipeline for this study.

### *4.1. MONAI Label, and 3D Slicer Integration*

MONAI Label is a free and open-source framework designed to streamline the development of AI-based applications aimed at reducing the time required to annotate radiology datasets (Diaz-Pinto et al., 2024). MONAI Label offers researchers the ability to develop AI annotation applications focusing on their domain of expertise. It allows researchers to readily deploy their apps as services, which can be made available to clinicians via their preferred user interface. Currently, MONAI Label has a plugin that can readily be integrated into the 3D Slicer application. This integration allows researchers to connect their AI architectures with 3D Slicer, providing both interactive and non-interactive segmentation capabilities. Clinicians can then correct segmentation results either manually or using the automatic AI correction features offered by DeepEdit and DeepGrow models. This significantly reduces the time radiologists need to annotate cases. For our research, the Monai Label integrated with 3D slicers is composed of three models, the segmentation model which can be used for automatic segmentation, the 3D DeepGrow model which can be used for automatic annotation and correction of our segmentation, and finally the DeepEdit model which can be used for both automatic segmentation and interactive correction.

#### *4.1.1. Segmentation Model*

This is the primary model for automatic segmentation of the organs which is embedded in the 3D slicer. The model was trained on 20 cases to offer an initial segmentation which we used as an inference to select unique cases based on the AL uncertainty strategy during the active learning phase. For this model, we utilized the UNETR (UNet TRansformers), which employs a transformer as the encoder to learn sequence representations of the input volume. This approach effectively captures global multi-scale information, while adhering to the successful ”U-shaped” network design for the encoder and decoder. The transformer encoder is directly connected to the decoder via skip connections at different resolutions, facilitating the computation of the final semantic segmentation output. This architecture was consistently used throughout the active learning stage. Figure 3 represents the architecture of UNETR as developed by the authors in (Hatamizadeh et al., 2022).

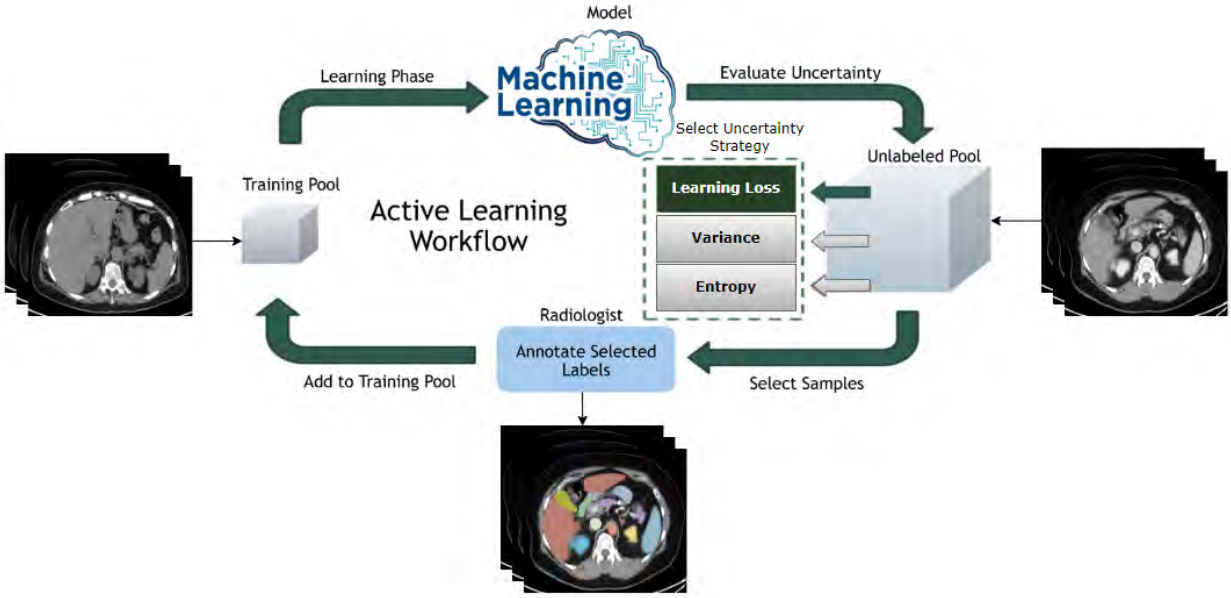



Figure 2: Full pipeline of our methodology

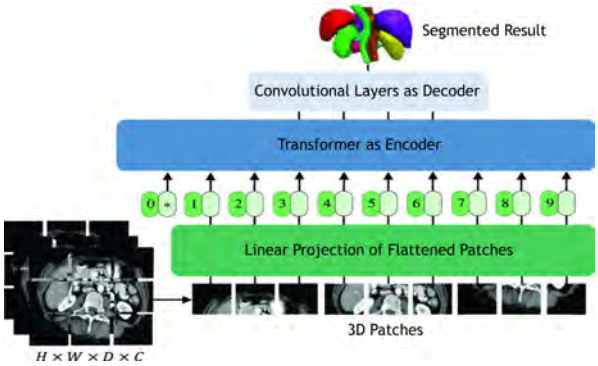



Figure 3: Overview of UNETR architecture (Hatamizadeh et al., 2022)

#### *4.1.2. 3D DeepGrow Model*

The 3D DeepGrow model is an interactive segmentation tool based on fully convolutional neural networks (FCNNs) (Long et al., 2015), where the user guides the segmentation process with positive and negative clicks. This model allows users to annotate one label at a time across the entire volume (Sakinis et al., 2019). Different organs or structures can be delineated based on the placement of these clicks and the selected label. The training process for a DeepGrow model differs from traditional deep learning segmentation due to the inclusion of positive and negative guidance (clicks) during training. Positive and negative guidance maps are generated based on false negatives and false positives, which depend on the predictions made by the model.

The aim of this model in our work is to simplify the process of annotating or correcting segmentation by automating it and enabling radiologists to refine the segmentation through interactive clicks. Positive clicks, known as foreground clicks, are placed within the structure of interest to guide the network in predicting the foreground (i.e., the organ to be segmented). If the segmentation result is under-segmented, additional foreground clicks can be placed in areas identified as false negatives to reduce these errors. Negative clicks, or background clicks, are placed where the current structure of interest is over-segmented to correct false positive errors, thereby guiding the network to reduce incorrect foreground predictions.

#### *4.1.3. DeepEdit Model*

The DeepEdit model merges the strengths of two approaches: automatic segmentation using dynamic U-Net (dynUNet), a variant of nnU-Net implemented in MONAI, and an interactive segmentation method from 3D DeepGrow, into a single deep learning model. More specifically, it allows the user to perform inference, as a standard segmentation method (i.e. dynUNet), and also to interactively segment part of an image using clicks as DeepGrow. The architecture of DeepEdit can utilize any segmentation network backbone, such as UNET (Ronneberger et al., 2015), nnU-Net (Isensee et al., 2019), UNETR (Hatamizadeh et al., 2022), SwinUNETR (Hatamizadeh et al., 2021), or dynUNet, allowing it to be tailored to specific needs.

It allows easy integration of uncertainty-based ranking strategies and active learning, making it a powerful tool for improving segmentation accuracy. The training process for DeepEdit involves two modes: standard training for automatic segmentation and training with user interaction simulation for interactive segmentation. During training, the input to the network is a concatenation of three tensors: the image itself, positive clicks

indicating the foreground, and negative clicks indicating the background. The training is divided into two stages. For half of the iterations, the tensors representing the foreground and background points are zeros, simulating a standard automatic segmentation model. For the other half, positive and negative clicks are simulated, training the model like 3D DeepGrow. This dual-stage training approach enables DeepEdit to perform fully automatic segmentation, semi-automatic segmentation with initial clicks, and refinement of existing segmentations through user-provided clicks.

Once trained, DeepEdit allows clinicians to efficiently segment datasets using either the auto-segmentation mode or by providing clicks via 3D Slicer. During inference, the radiologist can choose between automatic segmentation, where the click tensors are zeroed out, and interactive segmentation, where user-provided clicks guide the segmentation process. Figure 4 and Figure 5 represent the schematic representation of DeepEdit for both training and Inference mode. The difference in the figures is that clicks are simulated during training while in inference, it is interactive.

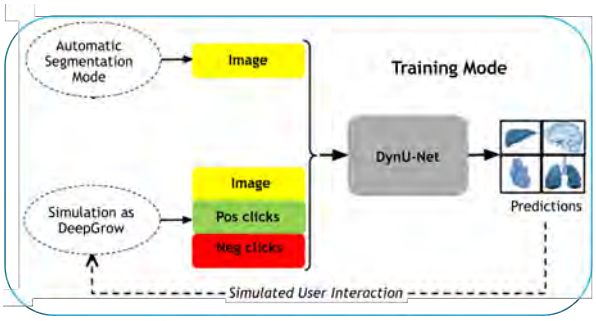



Figure 4: Schematic representation of DeepEdit during training mode.

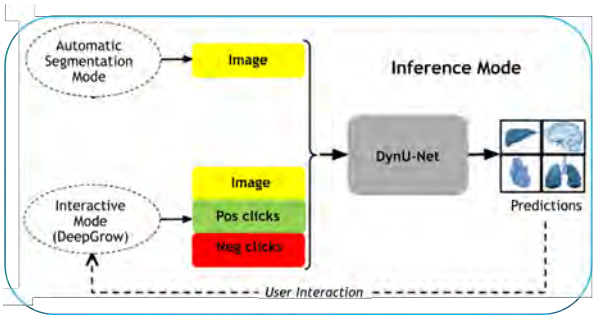



Figure 5: Schematic representation of DeepEdit during inference mode

### *4.2. Active Learning Implementation and Strategies*

Now that we have all models set up in an interactive working environment that supports active learning, we can go ahead and carry out the main task i.e. implementing several active learning strategies and testing them against the random data selection and comparing the results. For this work, we have implemented three active learning strategies, upon which two are based on uncertainty estimation with variance, or entropy using Monte Carlo Dropout, and the other is based on guiding the segmentation by strategically training a loss function and also using this trained function as an active learning strategy for data selection. The idea is to prove that AL strategies can be used to improve the performance of supervised models with a smaller number of annotated samples compared to random selection.

#### *4.2.1 Monte Carlo Dropout*

Monte Carlo Dropout is a technique where dropout layers, typically used during training to prevent overfitting, are also applied during inference. In our approach, we used a dropout rate of 0.2 in the UNETR architecture. We then performed 25 forward passes with different dropout masks, and generated a distribution of predictions for each input. This method allowed us to estimate the uncertainty of the model's predictions for two of the strategies we implemented: variance, and entropy. The variation across these predictions enabled us to identify and focus on the most uncertain samples for active learning.

#### *4.2.2. Uncertainty Estimation*

Uncertainty estimation in active learning involves determining which data points the model is least confident about. These data points are then prioritized for annotation, as their inclusion in the training set is expected to provide the most significant improvement in the model's performance. There are two primary types of uncertainties considered in active learning i.e. aleatoric uncertainty and epistemic uncertainty (Seoh, 2020).

Aleatoric uncertainty arises from the inherent noise in the data. It captures the variability within the data itself, which cannot be reduced even with more data. For example, low-quality images or ambiguous regions in medical scans contribute to aleatoric uncertainty. This type of uncertainty is intrinsic to the data and remains despite the quantity of data available.

Epistemic uncertainty, also known as model uncertainty, arises from the lack of knowledge about the model parameters. This uncertainty can be reduced by training the model with more data. Epistemic uncertainty is particularly high in regions where the model has not seen enough similar examples during training. Addressing this type of uncertainty is essential for improving the model's accuracy and robustness. Our active learning strategies i.e. variance, entropy, and learning loss focus on these uncertainties. By implementing these strategies, we can effectively identify and select the most uncertain samples for annotation, thereby improving the overall performance of the model.

#### *4.2.3. Variance*

We computed the variance as one of the ways to estimate the uncertainty of the model's predictions. This

process involves running multiple inferences on each input image using Monte Carlo Dropout, which enables us to capture the variability in the model's predictions. These predictions are then analyzed to compute the variance, which serves as an indicator of the model's uncertainty. Mathematically, the variance computation is performed by first obtaining a set of prediction probabilities for each voxel across the 25 forward passes of Monte Carlo Simulations. These predictions are accumulated and processed to calculate the variance. The variance for each voxel is computed using the following formula:

$$\text{Variance} = \sum_{i=1}^{N} \left( \frac{1}{N} \sum_{j=1}^{M} (x_{i,j} - \mu_i)^2 \right) \tag{1}$$

where $N$ is the number of forward passes (25 in our case), $M$ is the number of classes (excluding background), $x_{i,j}$ is the prediction probability for voxel $i$ in class $j$, and $\mu_i$ is the mean prediction probability for voxel $i$ across all forward passes.

After implementing this variance-based uncertainty estimation, we ranked the unlabeled data points based on their uncertainty scores. The most uncertain samples, identified by their high variance, are then selected for annotation. This approach ensures that the model learns from the most informative data, leading to faster and more efficient improvements in segmentation performance.

*4.2.4. Entropy*

To estimate the uncertainty of the model's predictions, we also computed the entropy. Entropy measures the uncertainty in the predicted probability distribution for each voxel, providing insights into how confident the model is about its predictions. The process involves running multiple inferences on each input image using Monte Carlo Dropout, allowing us to capture the variability in the model's predictions just as before. These predictions are then analyzed to compute the entropy, which serves as an indicator of the model's uncertainty. Mathematically, the entropy computation is performed after obtaining a set of prediction probabilities for each voxel via Monte Carlo Simulations. The mean probability for each class is computed, and the entropy for each voxel is then calculated with:

$$\text{Entropy} = \sum_{j=1}^{M} (p_{i,j} \log(p_{i,j} + \epsilon)) \tag{2}$$

where $M$ is the number of classes (excluding background), $p_{i,j}$ is the mean prediction probability for voxel $i$ in class $j$ across all forward passes, and $\epsilon$ is a small constant added to avoid the logarithm of zero.

The unlabeled data points was ranked based on their entropy scores, and the most uncertain samples, identified by their high entropy, were then selected for annotation.

*4.2.5 Learning Loss Function*

To implement the learning loss function, we trained the UNETR model along with a loss function. The rationale behind this is that we are not only using the learning loss as a criterion for selecting the most informative samples for annotation but also to guide the segmentation process. The loss function employs a global average pooling layer followed by fully connected layers to predict a single scalar value representing the loss for each input. The network's forward pass generates segmentation outputs, which are then utilized by the loss predictor to estimate the corresponding loss.

*4.2.5.1 Training with Learning Loss*

During training, the network is optimized using a combined loss function that includes both the segmentation loss and the loss prediction error. The segmentation loss is computed using the Dice loss function, which measures the overlap between the predicted and true segmentation. The predicted loss, obtained from the loss predictor, is compared against the actual segmentation loss using mean squared error (MSE). The total loss function is defined as:

$$\text{Total Loss} = \text{Segmentation Loss} + \lambda \cdot \text{Loss Prediction Error} \tag{3}$$

where $\lambda$ is a weighting factor that balances the contribution of the loss prediction error. In our case, we found 0.05 to be the optimal value for $\lambda$.

*4.2.5.2 Learning Loss as Active Learning Strategy*

For the active learning strategy, we use the predicted loss to rank and select the most informative samples from the unlabeled dataset. The process involves running multiple inferences with Monte Carlo Dropout enabled to obtain a distribution of predicted losses. The mean predicted loss is then computed for each sample, and the samples with the highest mean predicted loss are considered the most informative and are selected for annotation. Mathematically, the mean predicted loss for each sample is computed as:

$$\text{Mean Predicted Loss} = \frac{1}{N} \sum_{k=1}^{N} L^{(k)} \tag{4}$$

where $N$ is the number of Monte Carlo forward passes, and $L^{(k)}$ is the predicted loss for the $k$-th pass.

*4.3 Iterative Segmentation and Active Learning*

With a clear understanding of the models available on the platform, our approach utilizes the developed active learning strategies to select cases for segmentation in the 3D Slicer. The segmentation models perform the initial segmentation, and the results are corrected by the radiologist before proceeding to the next active learning iteration. We conducted a total of 7 iterations, and after each iteration, we added 10 of the most uncertain cases into the active learning process. This iterative process ensures that the model continuously improves by focusing on the most informative data, thereby enhancing the overall segmentation results. The model was trained using each of the strategies i.e. variance, entropy, and learning loss and the results were compared against the random selection method. As a whole, integrating interactive segmentation and label correction within each active learning iteration helped us not only refine the model's predictions but also leverage radiologist corrections to guide the learning process effectively.

## 5. Results

In this section, we present the results of our study, including the evaluation metrics, uterus segmentation results, and the outcomes for each active learning strategy implemented. The performance of each strategy i.e. variance-based, entropy-based, and learning loss is also analyzed and compared in this section to show the effectiveness of active learning in improving segmentation results.

*5.1 Evaluation Metrics*

The primary evaluation metric for this project is the Dice Score, which measures the similarity between the predicted segmentation and the ground truth segmentation. The Dice Score ranges from 0 to 1, where 0 indicates no overlap and 1 indicates perfect overlap. It is a widely used evaluation metric for segmentation tasks, particularly in medical imaging. Mathematically, the dice score is computed by:

$$\text{Dice Score} = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{5}$$

Where:

- $A$ is the set of voxels in the predicted segmentation.
- $B$ is the set of voxels in the ground truth segmentation.
- $|A \cap B|$ is the number of voxels that are common to both the predicted and ground truth segmentations (i.e., the intersection).
- $|A|$ is the number of voxels in the predicted segmentation.
- $|B|$ is the number of voxels in the ground truth segmentation.

This metric provides a balanced measure of both precision and recall, making it an effective metric for assessing the accuracy of our segmentation model during active learning. High dice scores indicate a good overlap between the predicted and ground truth segmentations, while low dice scores generally imply poor results.

*5.2 Uterus Segmentation Results*

We initially performed the uterus segmentation using the nnU-Net model to incorporate it into our set of labels. Out of the 50 annotated cases, 40 of these cases were used for training the nnU-Net model, and the remaining 10 cases were reserved for final testing to evaluate the model's performance, the model recorded an average dice score of 83% across all 10 cases. Figure 6 represents the visualization of the segmentation in pairs where the first image for each session shows the original CT scan without any segmentation overlay, providing a clear view of the anatomical structures, and the second image displays the same CT scans with the segmented uterus highlighted in blue.

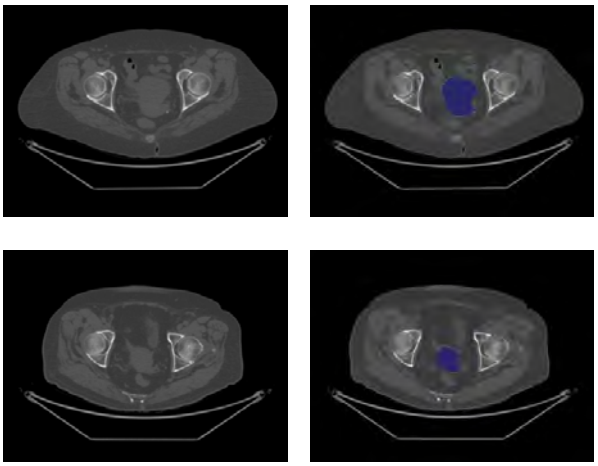


Figure 6: Results of the Uterus segmentation.

*5.3 Label Integration Results*

The integration of individual organ labels into a multi-class segmentation was crucial for our study. This process allowed us to consolidate 24 distinct organ labels into a single comprehensive dataset, enabling effective training of the active learning model on female cases that include the uterus. The integrated labels were visualized to assess their accuracy and effectiveness as shown in Figure 7. Additionally, Figure 8 showcases a 3D visualization of the labels highlighting the spatial

relationships between the various organs within the abdomen. The different colors in the 3D view represent different organs, showing the effectiveness of the label integration process in creating a detailed and accurate multi-class segmentation.

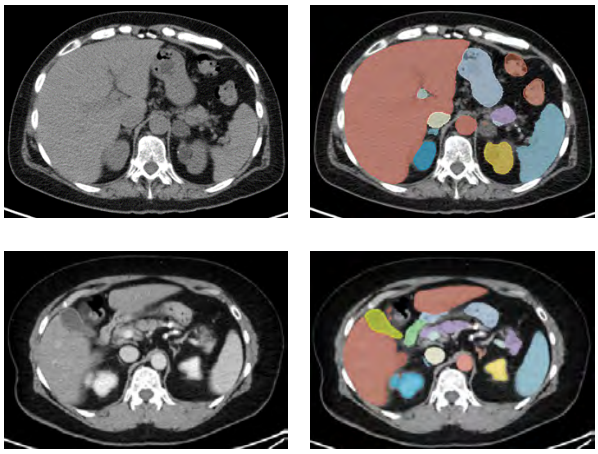


Figure 7: CT images with the integrated labels. The original image is on the left and the segmented organs are highlighted in different colors on the right.

Figure 8: Illustration of the 3D view of the integrated labels

*5.4 Baseline Results – Random Sampling*

The baseline for this work is based on random sampling, as the primary objective of active learning is to surpass the results achieved through random selection. This baseline provides a benchmark for evaluating the effectiveness of each active learning strategy. By comparing the results of various strategies to this baseline, we can determine the extent to which active learning enhances the performance of the segmentation model. In this strategy, data points are selected randomly from the pool of unlabeled data to be included in the training set. This process does not take into account any measure of uncertainty, making it a straightforward but essential point of reference.

*5.5 Radiologist Correction*

During the active learning process, the radiologist refines the model's predictions rather than annotating from scratch. To expedite this stage, we integrate multiple tools from MONAILabel, including the 3D DeepGrow model for automatic corrections and the manual segmentation editor within MONAILabel for precise adjustments. The advantage of the 3D DeepGrow tool lies in its ability to propagate corrections across all slices in a 3D volume, significantly reducing the time and effort needed for manual edits. Figure 9 illustrates the application of the 3D DeepGrow tool for correcting the segmentation of the liver and spleen. The top row displays the initial segmentation prediction, while the bottom row shows the refined segmentation after using the DeepGrow tool. The corrections are applied uniformly across the 3D volume, ensuring consistency in the final segmentation.

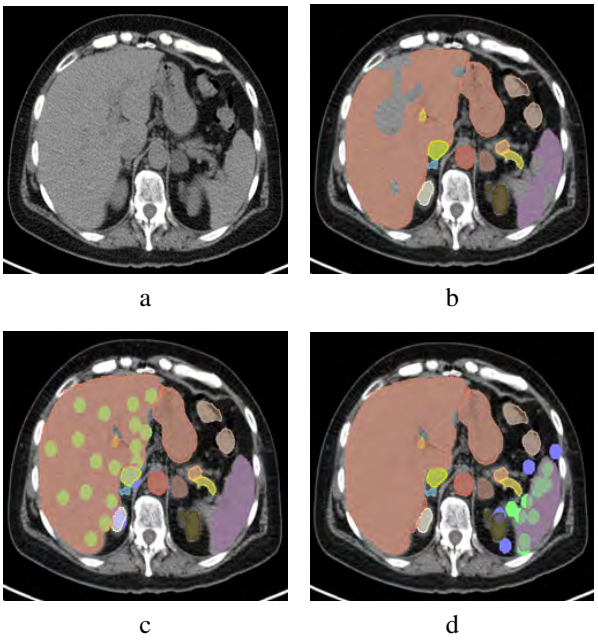



Figure 9: 3D DeepGrow tool for automatic correction of liver and spleen *(a. original slice, b. predicted slice, c. corrected prediction showing liver clicks, d. corrected prediction showing spleen clicks)*

*5.6 Impact of Active Learning Strategies*

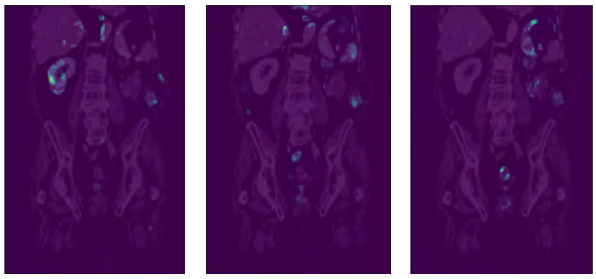
In this section, we present the results from the active learning strategy and explain how the model defines uncertainties by plotting the uncertainty maps.

*5.6.1 Variance-Based Strategy*

The variance-based strategy computes the variance to estimate the uncertainty in the model's predictions. By performing multiple forward passes using Monte Carlo Dropout, we captured the variability in the model's outputs. The variance is then calculated across these predictions, providing an indicator of uncertainty. The cases with the highest variance are selected to be in-

cluded in the training pool.

Figure 10 presents the predicted uncertainty map generated using variance. Areas with higher uncertainty are highlighted, indicating regions where the model is less confident in its predictions. Qualitatively, during iteration 1, the model displays significant uncertainty, particularly around major organs like the kidneys, while the liver exhibits less uncertainty. By iteration 4, the model has learned to identify the kidneys with greater certainty, reducing the uncertainty significantly. In the final iteration, the model demonstrates confidence in identifying both the kidneys and the liver, with minimal uncertainty remaining. This visualization illustrates the progression of the model's learning process, highlighting which parts of the image it initially finds ambiguous and where additional training data can be most beneficial when using the variance-based strategy.



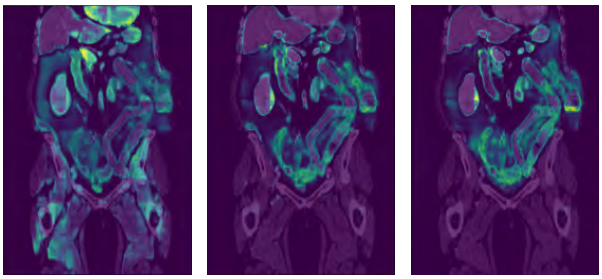
a. Iteration 1 b. Iteration 4 c. Iteration 7

Figure 10: Variance uncertainty map.

*5.6.2 Entropy-Based Strategy*

The entropy-based strategy calculates the entropy of the model's predictions to estimate uncertainty. In this method, the entropy for each voxel is calculated to quantify the uncertainty, and this value represents the amount of randomness or disorder in the predictions, with higher entropy indicating greater uncertainty. Cases with the highest entropy are selected to be included in the training pool.

Figure 11 demonstrates the predicted uncertainty map generated using entropy. Through this uncertainty map, we observed that organs that are easier to segment do not exhibit high entropy. For example, the liver is relatively the simplest organ to segment compared to others, and the figure illustrates that during iteration 1, the model shows significant uncertainty across many organs, including the liver. By iteration 4, the model becomes more certain about the liver, with very little uncertainty remaining in organs outside the abdominal region. By iteration 7, the uncertainty further diminishes, especially in the lower abdominal region, indicating that the model's predictions have become more refined and confident. This progression highlights that entropy as an uncertainty measure can provide valuable insights into the model's confidence at the pixel level and guide the selection of the most uncertain cases for training.



a. Iteration 1 b. Iteration 4 c. Iteration 7

Figure 11: Entropy uncertainty map.

*5.6.3 Learning Loss Strategy*

The learning loss strategy predicts the loss directly, providing a measure of the model's uncertainty about its predictions. This approach leverages a loss predictor, trained alongside the segmentation model, to estimate where the model's predictions are most likely to be incorrect. Incorporating this predicted loss into the active learning strategy allows the model to select data points where it is most uncertain, thus ensuring that the most informative samples are included in the training set.

Figure 17 shows the training curve comparing the random selection with and without the loss function guiding the segmentation. The results indicate that there are noticeable improvements when random selection is augmented with a loss function to guide the segmentation.

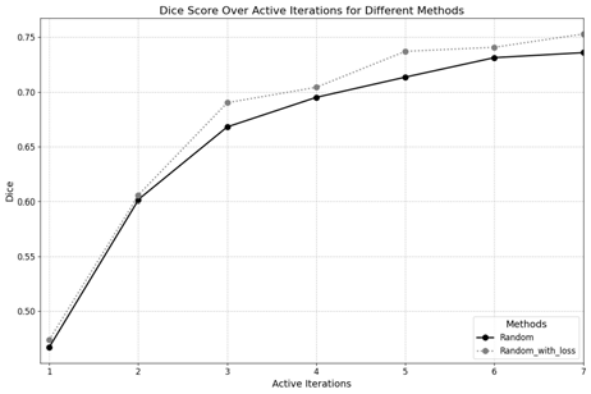



Figure 12: Iterative curve comparing random selection and learning loss-based strategy.

*5.6.4 Comparison of all strategies*

This comparison reveals that the learning loss strategy consistently outperforms all other methods, as illustrated in Figure 13. This implies that the method's ability to predict and leverage loss directly for guiding both segmentation and data selection provides a more focused learning process. The entropy and variance-based methods also demonstrate superior performance compared to random selection, indicating that these strategies can identify and prioritize the most uncertain data points for training.

The quantitative analysis presented in Table 2 shows the Dice scores of all implemented strategies during each iteration. From the data, we can observe that all active learning strategies, including random selection, show significant improvement in Dice scores from iteration 1 to iteration 2. This indicates that adding new training data in the initial iterations greatly enhances the model's performance. The learning loss strategy consistently achieves the highest Dice scores across all iterations. Starting from 0.5326 in the first iteration, it reaches 0.7770 by the eighth iteration. This demonstrates the strategy's effectiveness in continually improving the performance of the model.

Furthermore, both the variance and entropy-based strategies also show notable improvements over random selection. For example, in iteration 2, the Dice scores for variance and entropy are 0.6370 and 0.6307, respectively, compared to 0.6015 for random selection. However, their performance tends to plateau after the initial iterations, indicating that while they are effective, their impact diminishes as the model becomes more confident with increasing training data. As the iterations progress, the Dice scores for all strategies begin to converge, with less pronounced differences in the later stages. By iteration 7, the gap between random selection (0.7359) and the best-performing learning loss strategy (0.7770) is narrower compared to earlier iterations. This convergence suggests that the greatest benefits of active learning are realized in the early stages of training.

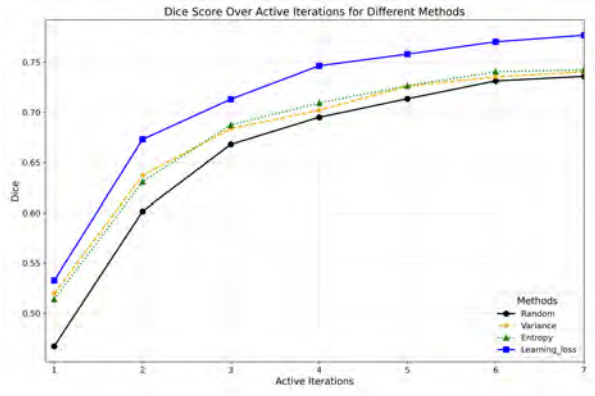



Figure 13: Result of active learning with all implemented strategies

Table 2: Dice scores of all implemented strategies during each iteration

| Iteration | Random | Variance | Entropy | Learning loss |
|---|---|---|---|---|
| 1 | 0.4670 | 0.5196 | 0.5140 | 0.5326 |
| 2 | 0.6015 | 0.6370 | 0.6307 | 0.6729 |
| 3 | 0.6680 | 0.6836 | 0.6875 | 0.7130 |
| 4 | 0.6950 | 0.7020 | 0.7093 | 0.7464 |
| 5 | 0.7135 | 0.7260 | 0.7265 | 0.7581 |
| 6 | 0.7313 | 0.7353 | 0.7405 | 0.7704 |
| 7 | 0.7359 | 0.7405 | 0.7424 | 0.7770 |

## 6. Discussion

From the results, as the iterations progress, all strategies (including random selection) converge towards similar Dice scores, highlighting that the initial iterations are crucial for maximizing the benefits of active learning. The variance and entropy-based methods, while effective, show a tendency to plateau sooner than the learning loss strategy. This suggests that while they are capable of identifying uncertain areas, their effectiveness may diminish as the model becomes more confident with increasing training data.

In contrast, the learning loss strategy's continued superiority suggests it provides a more robust mechanism for identifying areas where the model's predictions can be further refined. This consistency in the learning loss result can be attributed to its dual role in guiding the model during training and actively selecting the most challenging samples for annotation during active learning iterations.

Additionally, one of the key advantages of the learning loss strategy is its ability to predict loss directly, providing a direct measure of how uncertain the model is about its predictions while reducing the computational time. In contrast, other methods rely on indirect measures of uncertainty, such as variance in predictions or entropy of predicted probabilities, which might not always correlate perfectly with the model's actual performance on those samples and are computationally expensive. Also, the combination of segmentation with a loss prediction model allows the model not only to segment images but also to estimate its performance on the loss function, creating a self-assessment capability that enhances the efficiency of the learning process. This implies that the learning loss model can adapt to various datasets because it learns to predict loss based on the specific characteristics of the dataset during training.

Overall, the performance of the learning loss strategy continues to improve which means that it adapts dynamically as the model learns. Each iteration focuses on the current weaknesses of the model, ensuring that the learning process is always targeted at the most beneficial areas regardless of how confident the model is. This direct alignment of the learning loss strategy with the model's weaknesses ensures more targeted and effective learning, ultimately leading to sustained improvement over multiple iterations. The results demonstrate that integrating advanced active learning strategies, especially the learning loss strategy, can significantly enhance the performance of medical image segmentation models, and hence a powerful tool for improving diagnostic accuracy in clinical settings.

### *6.1 Limitation and Future Work*

While our studies have demonstrated that active learning strategies can significantly enhance segmentation models with fewer cases, several potential lim-

itations warrant further consideration. Active learning strategies depend heavily on the data, and their performance can vary significantly across different types of datasets. To ensure the robustness and generalizability of these methods, experiments should be conducted on a variety of datasets, including those with different imaging modalities (e.g., MRI, ultrasound) and anatomical regions using the learning loss strategy. By performing cross-validation and external validation on diverse datasets, the strategy can be fine-tuned to perform well across a broad range of scenarios. This approach will enhance the understanding and applicability of active learning in various medical imaging contexts.

In addition, the effectiveness of active learning methods is highly sensitive to the selection of hyperparameters, such as the dropout rate, the number of Monte Carlo samples, and the weight given to the loss prediction error. Conducting a series of experiments to systematically explore and optimize these hyperparameters can help in identifying the most effective configurations. Techniques like grid search, random search, or Bayesian optimization can be employed to find the optimal set of hyperparameters, though these methods could be computationally expensive. Hence, future studies should also consider developing efficient hyperparameter tuning techniques that balance computational cost and performance gains.

Computational complexity remains a key limitation of active learning including using entopy and varinace. Implementing Monte Carlo Dropout and performing multiple forward passes for variance and entropy calculations are computationally intensive (entropy and variance). Leveraging high-performance computing resources and parallel processing is essential in managing the computational load. To address this limitation, we implemented the learning loss which is faster, however further experiments can be conducted to explore other computational techniques, such as using approximation methods. Addressing these limitations involves further studies and experimentation focusing on hyperparameter tuning, improving generalization, and managing computational complexity.

## 7. Conclusions

We have demonstrated the potential of active learning strategies to enhance the performance of segmentation models in medical imaging. By systematically evaluating variance-based, entropy-based, and learning loss strategies, we found that these methods can significantly improve the efficiency of training segmentation models, particularly when labeled data is scarce. The learning loss strategy emerged as the most effective method, consistently outperforming other strategies, including random selection. This strategy's dual role in guiding the model during training and actively selecting the most challenging samples for annotation allowed for sustained improvement in segmentation performance. The variance and entropy-based methods also showed considerable performance, effectively identifying uncertain areas and prioritizing them for training, though their performance plateaued as the model became more confident.

Our findings highlight the importance of incorporating advanced active learning strategies into the training process of segmentation models. Doing this will not only improve the model performance but also enhance the training efficiency by focusing on the most informative samples. Future work should aim to validate these strategies across diverse datasets and imaging modalities, optimize hyperparameters systematically, and explore computational techniques to manage the complexity of active learning implementations.

## Acknowledgments

I would like to express my sincere gratitude to my supervisor, Dr. Joseph Lo, for his unwavering support and guidance throughout this project, as well as for providing me the opportunity to join the Center for Virtual Imaging Trials (CVIT) lab for my master's thesis. The lab environment and the warm welcome from all CVIT members were truly amazing, and I am deeply grateful for that. I also extend my heartfelt thanks to Dr. Mobina Ghojogh Nejad for her assistance with all the annotations, and to Lavsen Dahal for his valuable insights during the project. My deepest appreciation goes to the MAIA master program coordinators, the partner universities, and the European Union Commission for granting me the opportunity to pursue this exceptional program in medical imaging under the Erasmus+ scholarship. Lastly, I am profoundly thankful to my family and friends for their unwavering support, both physically and emotionally, throughout this journey.

## References

Arikan, M., Sallo, F., Montesel, A., Ahmed, H., Hagag, A., Book, M., Faatz, H., Cicinelli, M., Meshkinfamfard, S., Ongun, S., et al., 2023. Deep active learning for robust biomedical segmentation. bioRxiv , 2023–03.

Burmeister, J.M., Rosas, M.F., Hagemann, J., Kordt, J., Blum, J., Shabo, S., Bergner, B., Lippert, C., 2022. Less is more: A comparison of active learning strategies for 3d medical image segmentation. arXiv preprint arXiv:2207.00845 .

Chen, Q.Q., Sun, Z.H., Wei, C.F., Wu, E.Q., Ming, D., 2022. Semi-supervised 3d medical image segmentation based on dual-task consistent joint learning and task-level regularization. IEEE/ACM Transactions on Computational Biology and Bioinformatics .

Diaz-Pinto, A., Alle, S., Nath, V., Tang, Y., Ihsani, A., Asad, M., Pérez-García, F., Mehta, P., Li, W., Flores, M., et al., 2024. Monai label: A framework for ai-assisted interactive labeling of 3d medical images. Medical Image Analysis , 103207.

Ebrahimi, S., Gan, W., Chen, D., Biamby, G., Salahi, K., Laielli, M., Zhu, S., Darrell, T., 2020. Minimax active learning. arXiv preprint arXiv:2012.10467 .

Gaillochet, M., Desrosiers, C., Lombaert, H., 2023. Active learning for medical image segmentation with stochastic batches. Medical Image Analysis 90, 102958.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: international conference on machine learning, PMLR. pp. 1050–1059.

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

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

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D., 2021. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, in: International MICCAI Brainlesion Workshop, Springer. pp. 272–284.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D., 2022. Unetr: Transformers for 3d medical image segmentation, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 574–584.

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.

Isensee, F., Jäger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2019. Automated design of deep learning methods for biomedical image segmentation. arXiv preprint arXiv:1904.08128 .

Kim, W., Kanezaki, A., Tanaka, M., 2020. Unsupervised learning of image segmentation based on differentiable feature clustering. IEEE Transactions on Image Processing 29, 8055–8068.

Liang, X., Huang, D.S., 2018. Image segmentation fusion using weakly supervised trace-norm multi-task learning method. IET Image Processing 12, 1079–1085.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical image analysis 42, 60–88.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440.

Luo, X., Chen, J., Song, T., Wang, G., 2021. Semi-supervised medical image segmentation through dual-task consistency, in: Proceedings of the AAAI conference on artificial intelligence, pp. 8801–8809.

Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y., 2023. Segment anything model for medical image analysis: an experimental study. Medical Image Analysis 89, 102918.

Mouheb, K., Nejad, M.G., Dahal, L., Samei, E., Segars, W.P., Lo, J.Y., 2023. Large intestine 3d shape refinement using point diffusion models for digital phantom generation. arXiv preprint arXiv:2309.08289 .

Nath, V., Yang, D., Landman, B.A., Xu, D., Roth, H.R., 2020. Diminishing uncertainty within the training pool: Active learning for medical image segmentation. IEEE Transactions on Medical Imaging 40, 2534–2547.

Rajchl, M., Lee, M.C., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., Bai, W., Damodaram, M., Rutherford, M.A., Hajnal, J.V., Kainz, B., et al., 2016. Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. IEEE transactions on medical imaging 36, 674–683.

Ramesh, K., Kumar, G.K., Swapna, K., Datta, D., Rajest, S.S., 2021. A review of medical image segmentation algorithms. EAI Endorsed Transactions on Pervasive Health and Technology 7, e6–e6.

Razzak, M.I., Naz, S., Zaib, A., 2018. Deep learning for medical image processing: Overview, challenges and the future. Classification in BioApps: Automation of decision making , 323–350.

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.

Saidu, I.C., Csató, L., 2021. Active learning with bayesian unet for efficient semantic image segmentation. Journal of Imaging 7, 37.

Sakinis, T., Milletari, F., Roth, H., Korfiatis, P., Kostandy, P., Philbrick, K., Akkus, Z., Xu, Z., Xu, D., Erickson, B.J., 2019. Interactive segmentation of medical images through fully convolutional neural networks. arXiv preprint arXiv:1903.08205 .

Seoh, R., 2020. Qualitative analysis of monte carlo dropout. arXiv preprint arXiv:2007.01720 .

Shimizu, A., Kimoto, T., Kobatake, H., Nawano, S., Shinozaki, K., 2010. Automated pancreas segmentation from three-dimensional contrast-enhanced computed tomography. International journal of computer assisted radiology and surgery 5, 85–98.

Siddiqui, Y., Valentin, J., Nießner, M., 2020. Viewal: Active learning with viewpoint entropy for semantic segmentation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9433–9443.

Sourati, J., Gholipour, A., Dy, J.G., Kurugol, S., Warfield, S.K., 2018. Active deep learning with fisher information for patch-wise semantic segmentation, in: International Workshop on Deep Learning in Medical Image Analysis, Springer. pp. 83–91.

Tharwat, A., Schenck, W., 2023. A survey on active learning: state-of-the-art, practical challenges and research directions. Mathematics 11, 820.

Wang, G., Li, W., Zuluaga, M.A., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., et al., 2018a. Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE transactions on medical imaging 37, 1562–1573.

Wang, G., Zuluaga, M.A., Li, W., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., et al., 2018b. Deepigeos: a deep interactive geodesic framework for medical image segmentation. IEEE transactions on pattern analysis and machine intelligence 41, 1559–1572.

Wang, J., Chen, Z., Wang, L., Zhou, Q., 2019. An active learning with two-step query for medical image segmentation, in: 2019 International Conference on Medical Imaging Physics and Engineering (ICMIPE), IEEE. pp. 1–5.

Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., Nandi, A.K., 2022. Medical image segmentation using deep learning: A survey. IET Image Processing 16, 1243–1267.

Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al., 2023. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence 5.

Wolz, R., Chu, C., Misawa, K., Fujiwara, M., Mori, K., Rueckert, D., 2013. Automated abdominal multi-organ segmentation with subject-specific atlas generation. IEEE transactions on medical imaging 32, 1723–1730.

Wu, F., Marquez-Neila, P., Zheng, M., Rafii-Tari, H., Sznitman, R., 2024. Correlation-aware active learning for surgery video segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2010–2020.

Wu, J., Chen, J., Huang, D., 2022. Entropy-based active learning for object detection with progressive diversity constraint, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9397–9406.

Xiao, X., Lian, S., Luo, Z., Li, S., 2018. Weighted res-unet for high-quality retina vessel segmentation, in: 2018 9th international conference on information technology in medicine and education (ITME), IEEE. pp. 327–331.

Xie, J., Wu, Q., Zhu, R., 2024. Entropy-guided contrastive learning for semi-supervised medical image segmentation. IET Image Processing 18, 312–326.

Xu, Y., Tang, O., Tang, Y., Lee, H.H., Chen, Y., Gao, D., Han, S., Gao, R., Savona, M.R., Abramson, R.G., et al., 2020. Outlier guided optimization of abdominal segmentation, in: Medical Imaging 2020: Image Processing, SPIE. pp. 799–805.

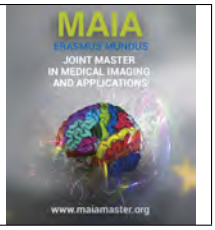
Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z., 2017. Suggestive annotation: A deep active learning framework for biomedical image segmentation, in: Medical Image Computing and Computer

Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20, Springer. pp. 399–407.

Zhao, Z., Zeng, Z., Xu, K., Chen, C., Guan, C., 2021. Dsal: Deeply supervised active learning from strong and weak labelers for biomedical image segmentation. IEEE journal of biomedical and health informatics 25, 3744–3751.

Medical Imaging and Applications

Master Thesis, June 2024




# Deep Learning-Aided End-to-End Uveitis Screening via Ultrasound Imaging

Yusuf Baran Tanrıverdi, Paula Petrone, Hassan Ahmed Sial

*Instituto de Salud Global de Barcelona, 08003 Barcelona*

**Abstract**

Uveitis is a significant cause of visual impairment, affecting 1.3 to 4.1 million people worldwide every year, making early and accessible screening crucial for reducing blindness. The gold-standard diagnosis involves counting white blood cells (WBC) in the anterior chamber using a Slit-Lamp examination, which is often inaccessible, difficult, and uncomfortable. Our novel Neosonics® transfontanellar ultrasound device offers a less invasive alternative. The diagnostic process can be further automated by deploying explainable-AI guided convolutional neural networks (CNNs) and decision mechanisms, easing the burden on clinicians. This study presents an end-to-end framework for uveitis screening that combines an innovative ultrasound technology with a three-stage deep-learning solution. This convolutional neural network (CNN) framework first applies quality control and segmentation of liquid in the anterior chamber (AC) of the eye. Next, a binary classification using a fine-tuned Resnet50 is deployed to detect white blood cells, indicating inflammation. A hard voting scheme is applied to decide the final diagnosis of anterior uveitis. Finally, explainable AI (xAI) techniques are deployed on top of the framework for further inspection. Our framework has shown promising results in low-resolution quality control with an accuracy of 85 %. Segmentation of cornea and liquid has reached Dice performance of 0.79 and 0.93, respectively. At the image level, the binary classification of cells achieved 90.83 % accuracy and 89.32 % F1 score. In a clinical slit-lamp-based study with 26 patients, our framework achieved a diagnostic accuracy of 25 out of 26 cases at the eye level, demonstrating its efficacy *in vivo*. The proposed framework shows potential in aiding the early detection and diagnosis of uveitis.

*Keywords:* Transfontanellar Ultrasound, Transfer Learning, Uveitis, Non-invasive Imaging, xAI

## 1. Introduction

Uveitis refers to a family of intraocular inflammatory conditions in the eye's middle layer. The worldwide incidence of uveitis is roughly in the range of 1.3 to 4.1 million new cases per year (Miserocchi et al., 2013) and is unfortunately responsible for vision loss in 5-20 % of patients in the USA & Europe (Bodaghi et al., 2001; ten Doesschate, 1982; Krumpaszky and Klauss, 1992) and probably accounts for 25 % of blindness in the developing world (Rothova et al., 1996). Numerous recent studies supported higher prevalence around the globe (Dandona et al., 2000; Seepongphun et al., 2021). Uveitis causes 2.8 - 10 % of cases of blindness in the working-age population (Darrell et al., 1962; Suttorp-Schulten and Rothova, 1996). Moreover, the trend of childhood uveitis has been reported to rise to 33 % (Päivönsalo-Hietanen et al., 2000). Children constitute 5-10 % of cases in tertiary centers (Edelsten et al., 2003). Pediatric uveitis poses a significant threat to vision, resulting in more than 20 % of the children experiencing vision loss in one or both eyes (Abd El Latif et al., 2019). Pistilli et al. (2021) highlighted risk factors for reduced vision, particularly among older individuals, Hispanics, and smokers, emphasizing the importance of specialized uveitis management. Al-Ani et al. (2020) showed moderate vision loss occurs in the follow-up period significantly, with a substantial portion resulting in permanent impairment. These findings underscore the need for specialized care and comprehensive management strategies in uveitis patients to mitigate visual complications.

Among all uveitis forms, Bodaghi et al. (2001) showcased that anterior uveitis (AU), where the anterior chamber (AC) has inflammation, was found to be the second most prevalent. The vast majority of children

with uveitis were diagnosed with AU (Eser-Ozturk and Sullu, 2020; Päivönsalo-Hietanen et al., 2000). Moreover, Al-Ani et al. (2020) showed chronic AU carries a great risk for vision loss. Despite its dominant contribution to vision loss globally (Rothova et al., 1996), it is not easy to diagnose AU. The current standard diagnostic procedure involves counting white blood cells (WBC) in AC by specialists using a slit lamp (Qian et al., 2021) and counting the WBC by expert visual inspection. However, this procedure introduces subjectivity and relies on the human eye, posing challenges to avoiding human errors and following treatment responses.

Efforts to define and standardize uveitis seriousness persist (Trusko et al., 2013). The standardization of the Uveitis Nomenclature (SUN) grading system, a scale of 0 to 4+, is now the conventional rule for clinical practice (Liu et al., 2020), yet challenges remain in achieving consistency. To secure the best clinical outcomes, ongoing reliable monitoring and involving more than one specialist in assessing the condition (Kempen et al., 2008) might be needed. The potential benefits of more quantitative, visual examination tools to provide objective assessments aim to reduce medical professionals' workload and consultation needs.

Patients and hospitals face high costs due to non-standardized follow-ups (de Parisot et al., 2020). This is further compounded by the challenge of securing experienced subspecialists trained in conventional imaging tools, such as slit-lamp.

In an ideal slit-lamp examination, light beams should project onto the anterior chamber precisely. The practitioner then focuses on whether the liquid in the anterior chamber appears completely black. If it does not, WBC will appear as tiny, flake-like white dots, increasing in number as the SUN grade rises. This complex and time-consuming process requires extensive training and expertise, making it difficult for many medical professionals to master (Deuchler et al., 2023; Kaur and Gurnani, 2024). Even with this level of mastery, the diagnosis and grading of AU are highly subjective (Jabs et al., 2018; Konstantopoulou et al., 2012) and more unreliable than instrument-based technologies or automatized tools (Liu et al., 2020).

Although there are alternatives such as optical coherence tomography (OCT) and confocal microscopy, these technologies are often inaccessible to the target patient group for uveitis, particularly in developing or underserved regions (Jennings et al., 2022; Okonkwo et al., 2023). Moreover, their practical use is subject to different protocols, and it is often hard to distinguish inflammatory cells from other cell types smoothly (Maring et al., 2022; St. Croix et al., 2005). In addition, all these alternatives require patients to remain stationary and keep their eyes open for long examinations. This introduces discomfort and difficulty in screening, especially for children. Additionally, the World Health Organization (WHO) has reported the lack of affordability for medical imaging in about two-thirds of the world (Anderson et al., 2003; Bélard et al., 2016). This often prompts medical sectors to use more affordable and portable tools such as ultrasound (US) to address a giant global logistical and inaccessibility problem (Dietrich et al., 2019).

There have been recent advances for the US in uveitis detection and were solely targeted for emergency and primary care points (Ortiz-González et al., 2024; Tabbut et al., 2019; Zur et al., 2016). However, to realize its full potential, ultrasound should not be regarded solely as an imaging modality, but also as an interactively integrated end-to-end clinical assessment tool (Dietrich et al., 2019; Rix et al., 2018). Artificial intelligence (AI), specifically convolutional neural networks (CNN) can aid in building such a framework and enhance US's performance in quantitative and preferably automatized tasks.

Given its deeper penetration capability, usability, and non-invasiveness, ultrasound (US) has significant potential as a screening tool compared to its competitors: slit-lamp, confocal microscopy, and OCT. The Neosonics® device, a cutting-edge, non-invasive ultrasound screening tool, has been used to detect backscattering signals from WBC in body fluids. The device designed for high resolution aims to surpass the current diagnostic capabilities of ultrasound technology, focusing on detecting inflammation in the CSF that arises due to meningitis in newborns (Ajanovic et al., 2023). Recent research has shown that automatic cell count-to-grade diagnosis is possible for 2D US images (Ajanovic et al., 2023; Sial et al., 2024), utilizing a common CNN architecture, Resnet50 (He et al., 2016).

Explainable AI (xAI) techniques, which provide transparency in deep learning models, are gaining research impact and are considered more ethical, especially in healthcare applications. Additionally, xAI and high interpretability enhance the understanding of experts and patients by revealing AI's 'black-box' nature' (Amann et al., 2020; Arrieta et al., 2020; Cinà et al., 2022; Srinivasu et al., 2022). This approach is in line with modern AI transparency standards and is crucial for our work in infant meningitis screening. Understanding and visualizing the key patterns from WBC backscatter signals in the anterior chamber liquid are critical for advancing the diagnostic process and improving patient care.

In this paper, we deploy a pipeline to achieve end-to-end automatic screening of Uveitis, and the detection of WBC in AC *in vivo*. The first stage includes a quality control check of processed low-resolution (LR) images, i.e., using the device in non-focal mode and identifying eye structures in the middle layer, such as cornea and liquid. In the second stage, high resolution (HR) deep-learning binary classification of images identifies the presence of WBC to detect the disease at the image

level. Finally, a hard voting scheme is applied to integrate the information of a set of eye images to decide the final diagnosis at eye level. In the last stage, we address the explainability and interpretability of the framework. The main contributions of this work include:

1. *Proof-of-concept:* Presenting a clinic-friendly explainable deep learning framework on *in vivo*, i.e., human eyes,
2. *End-to-End Application:* Building an end-to-end diagnostic screening methodology within Neosonics® US device,
3. *Standardization:* Enhancing overall screening performance of anterior uveitis integrating info from several images to ensure maximum performance.

## 2. State of the art

Artificial Intelligence (AI), particularly, deep learning techniques, has transformed medical diagnostics, including US imaging. This breakthrough has reached a level where it is now comparable to human-level analysis in medical imaging (Lee et al., 2017). At the center of these advancements, CNN models such as Resnet50 reside and are often specialized for classification, segmentation, or detection tasks (Krichen, 2023). Medical domains utilizing US imaging also greatly benefit from CNN-based technology (Yi et al., 2021). Its incorporation in US imaging has shown promise due to fast screening and diagnostic insights, which benefit from real-time algorithm deployment (Ajanovic et al., 2023; Sial et al., 2024). The main drawbacks of the US, e.g., clinically confusing scattering noise and artifacts (Raheem, 2021; Wu et al., 2020), might be addressed efficiently by enhancing the US diagnostic capabilities (Dietrich et al., 2019).

Ultrasound combined with AI has enabled accurate diagnoses of various medical conditions. Lung ultrasound helped diagnose COVID-19 and pneumonia (Buda et al., 2020). 3D transrectal ultrasound and deep learning were used for prostate segmentation (Orlando et al., 2020). CNN-based techniques identified cancerous tumors (Chi et al., 2017; Liu et al., 2017), and intelligent detection tools facilitated breast tumor detection (Zhang et al., 2020). Transfer learning addressed data scarcity issues, enhancing performance in liver fibrosis classification (Meng et al., 2017; Yi et al., 2021). AI-powered ultrasound diagnostics graded inflammation severity (Lin et al., 2020), detected early gastrointestinal inflammation (Yang et al., 2021), classified inflammatory myopathies (Uçar, 2022), and monitored fatty liver severity (Chou et al., 2021).

In light of these recent advancements, eye care was also revolutionized. Wang et al. (2021b) and Zhang et al. (2020) utilized B-ultrasound images to detect cataracts using feature extraction and deep learning. Others used ultrasound biomicroscopy to localize eye structures and automatically assess AC angle (Shi et al., 2019; Wang et al., 2021a). Various US applications for detecting uveitis or inflammation in the eye also exist (Häring et al., 1998; Zur et al., 2016) and are often designated for fast primary screening and urgent treatment (Hoffmann et al., 2020; Ortiz-González et al., 2024; Tabbut et al., 2019). Although found valuable for diagnostics (Fledelius, 1996), none have developed a completely automated solution using AI as far as our knowledge reaches.

Fortuitously, when addressing the challenge of examining 'inflamed cells' within bodily fluids, particularly the aqueous humor (liquid of AC) in our case, innovative methodologies such as Neosonics® transfontanellar ultrasound (Ajanovic et al., 2023) have emerged to alleviate concerns regarding inter-observer variability. This noninvasive portable and accessible device utilizes backscattering signals from tissues. The backscattering data from the US was proven effective in estimating the concentration of various cell suspensions *in vitro* settings (Elvira et al., 2023; Jimenez et al., 2016; Lee et al., 2018). Moreover, Ajanovic et al. (2023) and Sial et al. (2024) classified Meningitis and control groups of infants using this technology *in vivo*. These methodologies aim to enhance visualization, streamline the diagnostic process, and consequently facilitate automation using computational intelligence.

In this paper, we demonstrate the *in vivo* application of this technology for eye care, specifically for anterior uveitis, showcasing its potential to revolutionize diagnostic approaches in this field.

## 3. Material and methods

### *3.1. Dataset*

#### *3.1.1. Data Acquisition*

We use data recruited *in vivo* across *Hospital Germans Trias i Pujol* in Barcelona, Spain. Ultrasound screening was performed on each patient's left and right eye using the Neosonics® device, while their eyelids were closed and stretched with a piece of tape. The Neosonics® device scans the eye top-to-bottom and collects 2D images of the eye. At the end, images from all different scans of the eye compose its eye folder. The initial low resolution dataset has 9781 images from 447 eyes. After a series of quality controls at both frame and eye levels, a total dataset of 1069 HR 2D ultrasound images from 26 eyes was used. These quality controls include the low-resolution stage and post-clinical decisions made by our technicians.

The proposed framework uses existing data in a top-down approach. Stage 1 begins with the largest database (N=9781), where a substantial amount of experimental data is available. In this stage, many scans were intentionally performed incorrectly and randomly to create quality-control "bad" cases. However, only those

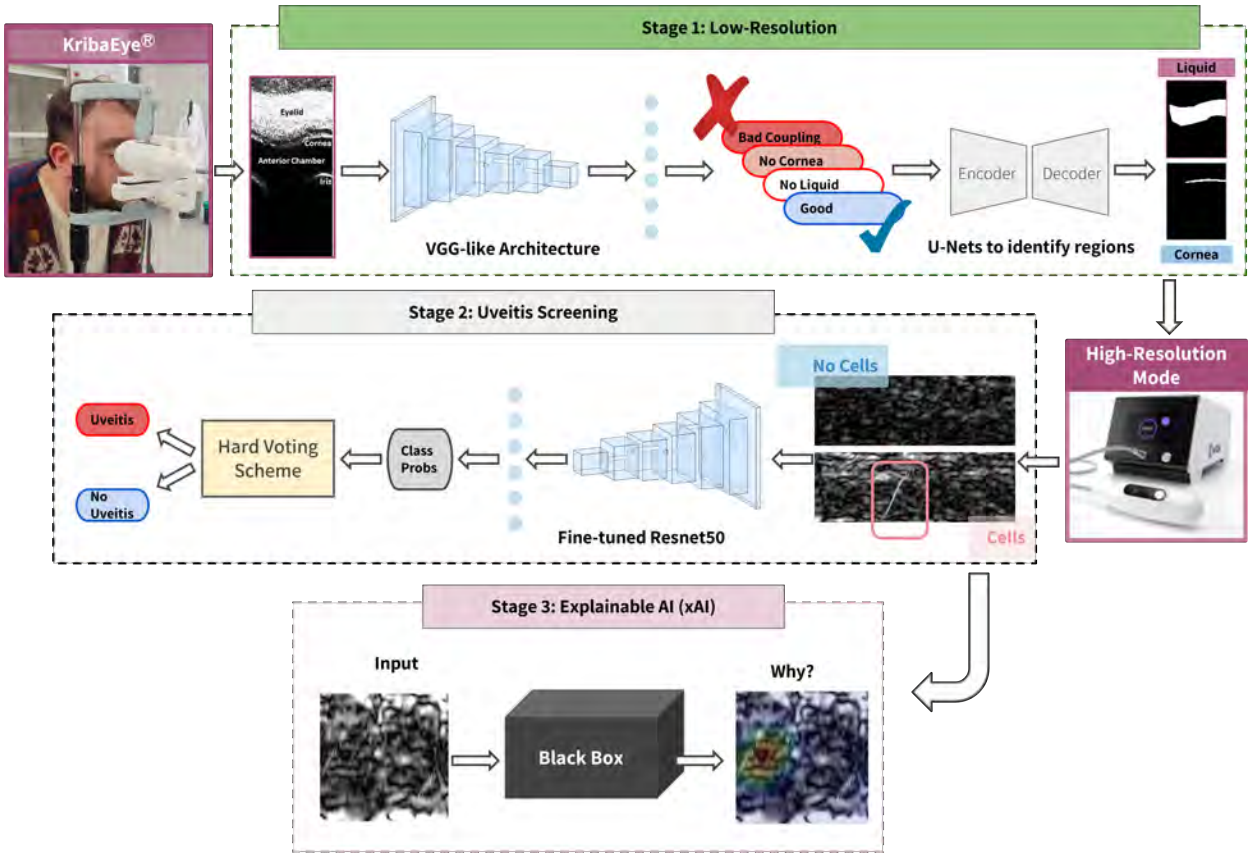



Figure 1: Flowchart of the three-stage methodology employed for anterior uveitis diagnostic screening using ultrasound images. Images are acquired on the patient's eyelid using novel Neosonics® technology. Stage 1 (Quality Control and Segmentation): VGG-like deep learning architecture filters out ultrasound images that exhibit bad quality. Then, U-Net is used to segment the cornea and liquid. Stage 2 (Screening) employs binary classification to distinguish between "Cells" and "No Cells" images based on the presence of increased WBC cellularity in the AC as visualized in ultrasound images and applies hard voting to decide the final diagnosis. Stage 3 (Explainable Artificial Intelligence (xAI): xAI techniques are applied for model interpretability.

with "good" labels and manually annotated images were used for segmentation. Subsequently, the best area selection algorithm is tested on these successfully segmented images internally, paving the way for transitioning to the high-resolution (HR) setting of the device.

In Stage 2 Uveitis Screening, the selection of eyes from patients for inclusion in the study adhered to specific criteria. These criteria primarily focused on the clinical and technological aspects of inclusion. Clinically, emphasis was placed on the absence of abnormal eye structures and the patient's ability to undergo the slit lamp test administered by a doctor. Technologically, cases were excluded if the data lacked sufficient HR quality due to issues such as excessive noise, movement, or attenuation. Finally, only 26 eyes from 20 patients were used in Stage 2. Anatomically, a person's left and right eyes may have different uveitis diagnoses and slightly varying structures. Therefore, in this study, we refer to each eye as a separate patient case, contrary to the typical individual perspective.

### *3.1.2. NeoSonics® Device*

Neosonics® is a cutting-edge, non-invasive ultrasound technology designed to detect backscatter signals from white blood cells (WBC) in the anterior chamber, which is the space inside the eye found between the cornea's inner surface and the iris. The Neosonics® device is a novel non-commercially available device that has not yet received any certification clearance. It makes spatial scans with steps smaller than 5 microns, which allows capturing the backscatter signals of individual cells within the liquid of AC, crucial for analyzing the composition of serous body fluids in a non-invasive way and at high sensitivity to structural changes not captured by conventional ultrasound systems (Ajanovic et al., 2023). For ultrasound imaging data collection, we used the Neosonics® ultrasound probe positioned over the closed or stretched eyelids of the patients. Figure 2 shows a mock examination and the device.

### *3.1.3. Ground Truth Generation*

*Low-Resolution.* Figure 3 illustrates labels used as ground truth for LRQC classification. These labels were

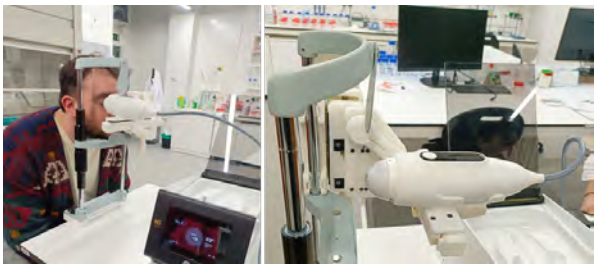


Figure 2: Pictures from our company trip. (left) The Author in mock-examination (right) Neosonics® US device.

annotated according to the following criteria.

- Bad coupling: Noise, device badly placed or other coupling artifacts are present.
- No Liquid: Good coupling but the liquid is absent in the image.
- No Cornea: Good coupling but the cornea is absent or not well defined.
- Good: The overall image quality is adequate; liquid and cornea are defined.

Moreover, cornea and liquid masks were drawn manually using basic computer skills. These ground truths were generated by the team from KRIBA.AI and double-checked by at least two people.

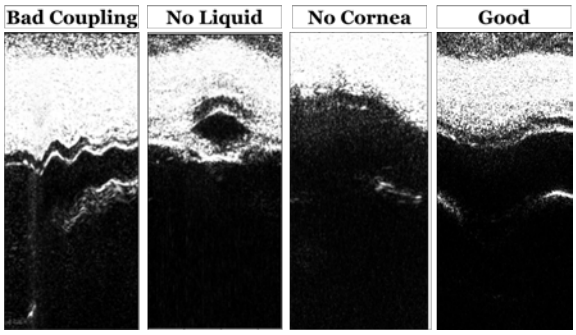



Figure 3: LRQC ground truths representing Bad Coupling, No Liquid, No Cornea, and Good, left to right respectively. Note that the proposed algorithm seeks both the cornea and liquid area for the next step, Low-Resolution Segmentation.

*Uveitis Screening.* SUN-grading-based ground truth was used for the dataset. The SUN grading system is an effort to standardize Uveitis diagnosis, relying on the count of WBC detected in the anterior chamber. Table 1 depicts each SUN grade's corresponding cell amount estimated in liquid (Chang et al., 2008). Considering the SUN grading scheme, the clinical threshold for a uveitis diagnosis starts at SUN 0.5. SUN reflects the severity of inflammation, proportional to the number of cells that reside in the anterior chamber. In the gold-standard diagnosis of anterior uveitis, using a slit-lamp,

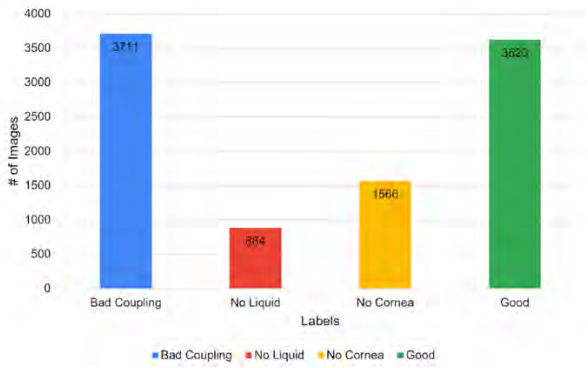



Figure 4: Bar charts showing the amount of images per LRQC label.

the SUN grade increases as the clinician observes more cells. Parallel to slit lamp examination, our Neosonics® device can also visualize WBC in this manner, visualized in Figure 5. However, both of the tools have challenges when it comes to SUN 0.5 grade (Konstantopoulou et al., 2012). This severity of uveitis implies there is a chance that cells will not appear in the liquid. Quantitative screening such as slit-lamp results in many patients not being diagnosed correctly, i.e., a high false negative rate. In the proposed quantitative screening with ultrasound, clinicians can ensure maximal sensitivity by taking multiple scans and integrating an intelligent framework. Therefore, each frame was annotated according to the cells' visual presence in our proposed framework. These annotations were done by the team from KRIBA.AI and double-checked by at least two people.

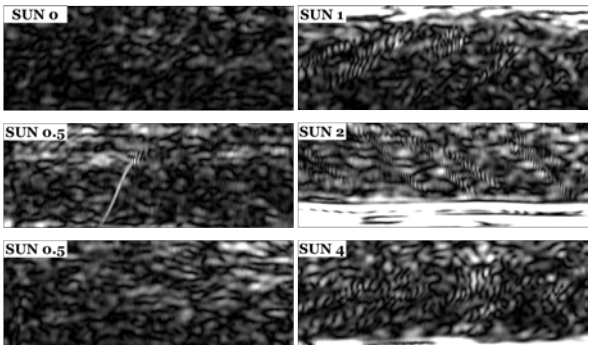



Figure 5: The SUN grades' visualization by Neosonics® device. Note that the number of cells increases as the SUN grade increases. In SUN 0.5, there is a probability of cells not appearing, necessitating multiple scans of the same eye while examining.

Each eye was graded by the Hospital Universitari Germans Trias i Pujol (HUGTiP) ophthalmological unit using a slit-lamp examination, in which the medical specialist counts the cell with the slit lamp and grades the eye using SUN. This generated SUN ground truth and was checked by one medical specialist. Figure 6 shows the distribution of SUN grades across our dataset.

| Grade | Cells in Field | AC Flare | Concentration (cell/µl) | Presence of WBCs |
|---|---|---|---|---|
| 0 | <1 | None | 0 | None |
| 0.5 | 1–5 | - | 1-50 | **Possible** |
| 1 | 6–15 | Faint | 51-150 | Definite |
| 2 | 16–25 | Moderate | 151-250 | Definite |
| 3 | 26–50 | Marked | 251-500 | Definite |
| 4 | >50 | Intense | >500 | Definite |

Table 1: The SUN Working Group Grading Scheme for Anterior Chamber Cells. Note that the field size is a 1 mm by 1 mm slit beam.

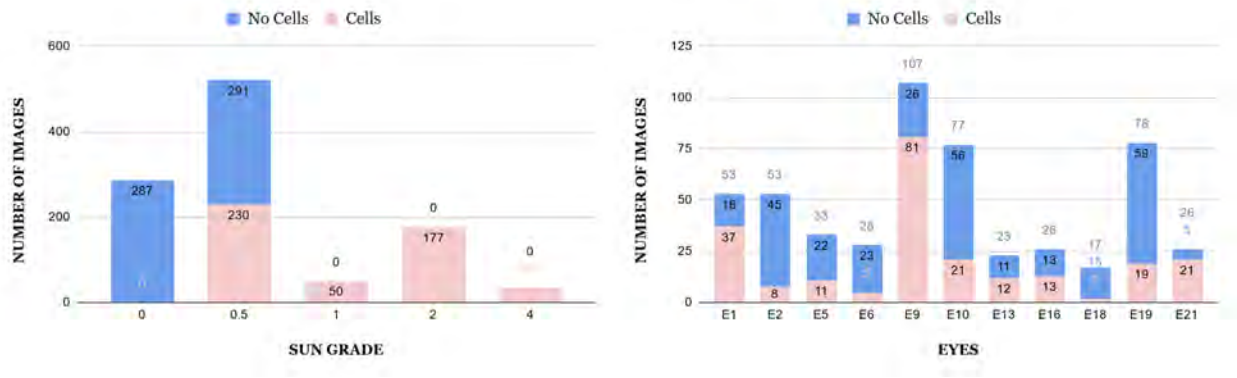



Figure 6: Bar charts of (left) image level count of SUN grades and (right) images with cell appearance of SUN 0.5 graded eyes.

### *3.1.4. Preprocessing*

Initial ultrasound signals acquired from the eye represent each vertical row scanned. These row signals are first denoised with a Butterworth filter, a conventional bandpass filter designed to allow frequencies between specified low-cut and high-cut values to pass while attenuating frequencies outside this range. Next, the Hilbert transform is used to extract the final shape of the ultrasound signal. After integrating the individual scans into 2D images, an interval threshold is applied to eliminate spikes. Finally, 2D images are normalized to 8-bit to appeal visually.

### *3.2. Pipeline*

Figure 1 shows our pipeline in this study.

### *3.2.1. Stage 1. Low-Resolution*

Low-resolution images can identify relevant eye structures: eyelid, cornea, and lens. However, these images may not provide useful information on white blood cells for two reasons: i) They might need some of the structures, i.e., bad quality, and ii) they need to focus on AC. Therefore, an internal algorithm needs to be deployed to the Neosonics® device. This algorithm would be responsible for

1. Quality control at the image level,
2. Segmentation of eye structures, cornea, and liquid for each frame acquitted in the scan,
3. Selecting frames fulfilling the criteria of having enough "liquid" area,
4. Selecting the best area coordinates within the cornea and liquid masks,
5. Proceeding HR mode concerning these coordinates.

Figure 7 illustrates this process clearly. The device focuses on low-resolution structural details and identifies and recommends the best region to be zoomed in. The operator accepts the recommendation and a new zoom-in acquisition is done around the focal area in high resolution. Finally, that area is cropped according to the internal purely mathematical algorithm.

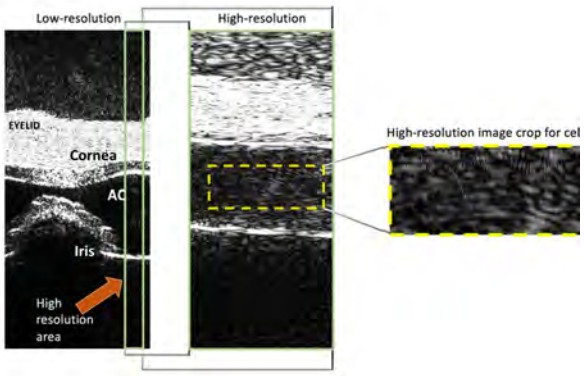



Figure 7: Depiction of low-resolution to high-resolution switch. It is seen that the Anterior Chamber (AC) (left) should be segmented to decide the focal area, which will be further cropped for cell counting (right).

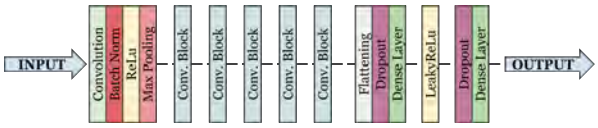



Figure 8: VGG-like custom CNN architecture. We use Dropout with a 0.5 rate and utilize Batch Normalization to prevent overfitting.

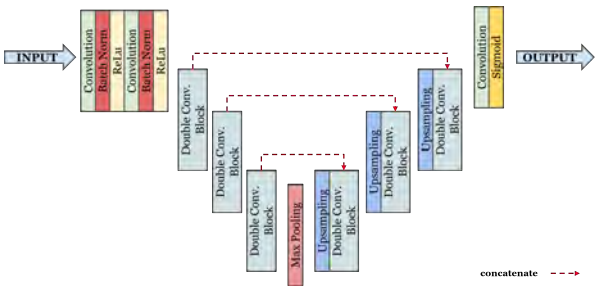



Figure 9: U-Net architecture.

*Low-Resolution Quality Control (LRQC).* Quality control is necessary before switching HR to ensure an efficient HR entry point. For this purpose, a larger dataset was built and annotated with four labels. To accomplish LRQC classification, a VGG-like architecture was deployed. The model consists of 6 convolutional layers that include batch normalization and LeakyRelu activation. Then, a classification head with 3 fully connected (FC) layers was deployed. A dropout functionality followed each FC layer with a rate of 0.5. All these choices were made to avoid overfitting in validation. The model's architecture is illustrated in Figure 8.

*Low-Resolution Segmentation of Cornea and Liquid.* The main structures to define the target area of AU are the cornea and liquid. WBC can be found in AC liquid or between AC liquid and cornea. Therefore, both are considered for the best area selection algorithm of the device. Two randomly initialized U-Net, a renowned CNN architecture (Ronneberger et al., 2015), were trained to specialize in cornea and liquid segmentation. A 4-depth variation of U-Net with double convolutional blocks was adopted. The architecture of U-Net is shown in Figure 9. The selection of the best area for the positioning of the acoustic beam focus in the acquisition's high-resolution phase is based on a calculation using both predicted cornea and liquid masks obtained from the segmentation model. The algorithm attempts to find the area inside the liquid that is found below the cornea, as the probability of finding WBC at this point is higher because of the signal transmission. This computation ensures being inside the predicted liquid area and under the predicted cornea.

### 3.2.2. *Stage 2. Uveitis Screening*

Figure 10 describes our workflow at this stage.

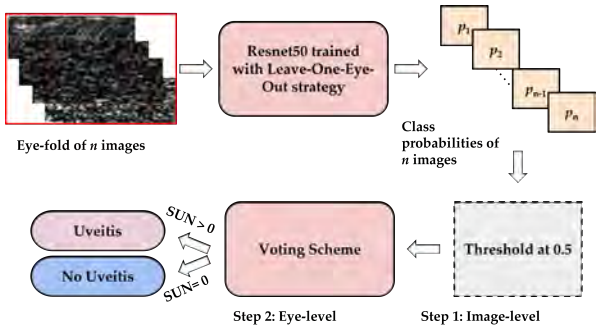



Figure 10: Flowchart of the diagnosis methodology using high-resolution ultrasound images. Step 1 (Image Level): pretrained Resnet50 is fine-tuned with the Leave-One-Eye-Out strategy, i.e., the rest of the eyes were in the training set. Probabilities are counted as predictions to distinguish between "Cells" and "No Cells" images according to a cut-off threshold of 0.5. This classification is based on the model's confidence in seeing an increased WBC cellularity. Step 2 (Eye Level): Hard and soft voting schemes were applied to test recall and accuracy performance.

*High-Resolution Binary Classification.* After acquiring high-resolution images, our proposed framework identifies WBCs within the frames using a transfer learning-based deep learning algorithm. This approach benefits pre-trained weights from models like ResNet50 (He et al., 2016), which are periodically updated using Python libraries such as PyTorch. These weights, originally trained on the ImageNet dataset (Deng et al., 2009), accelerate convergence by avoiding random initialization and effectively learning basic image features like edges. In our implementation, we modify the ResNet50 model's fully connected layer to have a single output channel for class probability. Figure 11 describes the model architecture.

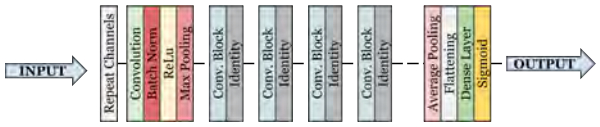



Figure 11: ResNet50 model architecture. The only modification is to have a single output channel for binary classification.

*Training Strategy.* In our study, we utilize a dataset comprising images from the same eye, which naturally exhibits similar anatomical structures. To maintain the integrity and validity of our model, it is essential to prevent splitting images from the same eye between the training and validation sets. Such a split could result in data leakage, leading to the model being trained on information overly similar to the validation data, thus artificially inflating performance metrics. This does not accurately reflect the model's real-world performance in an end-to-end application, where it is expected to generalize to entirely unseen patient data. To address this, we implemented a Leave-One-Eye-Out strategy, where all images from a given patient or group are ex-

clusively included in the training or validation set. This method ensures that the validation set remains independent, providing a more realistic and reliable assessment of the model's generalizability and performance in clinical practice. We utilized the built-in function for the leave-one-group-out split method in Python's *sklearn* library.

*Eye-Level Diagnosis.* With class probabilities as output from ResNet50, a decision mechanism is necessary to determine the final diagnosis: Is the SUN grade greater than 0? This study reported two different schemes—hard voting and soft voting—. Hard voting relies on image-level predictions, requiring a certain percentage of images to be classified as 'Cells' to diagnose uveitis. Conversely, soft voting considers the average of probabilities, diagnosing uveitis if the probability exceeds a specific cut-off threshold. Hard voting was found favorable in terms of recall performance within our dataset.

### 3.2.3. Stage 3. Explainable AI (xAI)

*GradCAM.* We employ GradCAM (Selvaraju et al., 2020) to visualize and scrutinize the model's learned attributes within images for identifying WBCs. GradCAM (Gradient-weighted Class Activation Mapping) accentuates significant regions in the image where the model concentrates, aiding in comprehending its predictive process. This visualization helps in understanding how the model identifies white blood cells and ensures that it focuses on relevant features.

Figure 12 illustrates the GradCAM flow. The process starts by computing the gradient of the score for a class concerning the feature maps of the last convolutional layer of Resnet50. These gradients are then globally averaged to obtain the importance weights for each feature map. The importance weights are used to perform a weighted combination of the forward activation maps. The resulting map is passed through a ReLU activation function to discard negative values. This GradCAM map is finally superimposed onto the original input, providing a visual representation of the areas the model focuses on for making its predictions.

By overlaying the GradCAM map on the input image, we can visually inspect which patterns in the image are most influential in the model's decision-making process. This helps in verifying the model's focus on pertinent features such as the morphology and structure of white blood cells, thereby enhancing the interpretability and trustworthiness of the model.

*UMAP.* UMAP (Uniform Manifold Approximation and Projection) is a dimensionality reduction technique that preserves global and local data structure when projecting from higher to lower dimensions (McInnes et al., 2018). It typically involves two steps: first, computing a graph that represents the data, and second, learning an embedding for that graph using non-parametric

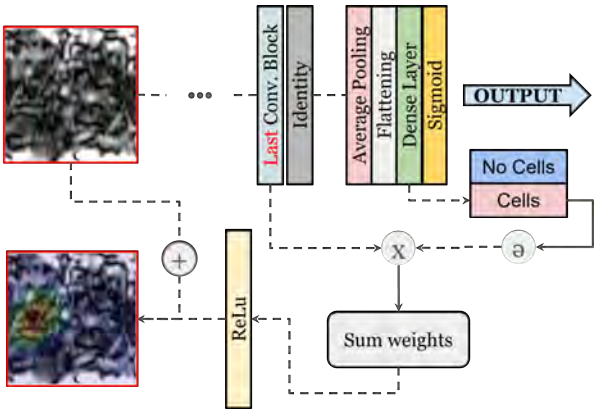



Figure 12: GradCAM flowchart.

clustering via UMAP. These steps facilitate a smooth reduction, reflecting how easily or difficult raw data can be clustered. Additionally, the second step can be replaced by a parametric approach, where the relationship between the data and neural network feature maps is learned (Sainburg et al., 2021). This makes UMAP especially valuable for interpreting neural networks, as it helps researchers visualize the complex, high-dimensional outputs of these networks in a clearer, two- or three-dimensional form.

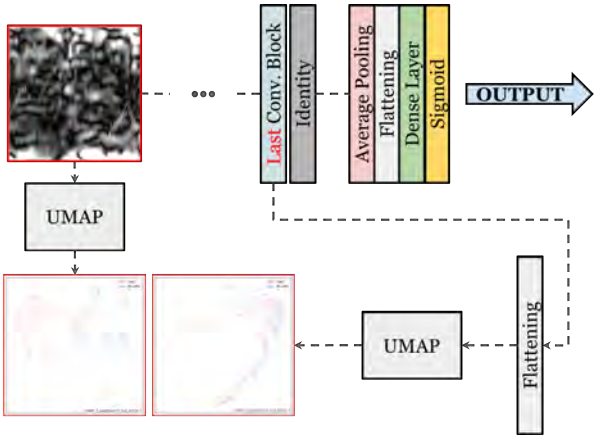



Figure 13: UMAP flowchart.

Figure 13 illustrates UMAP's role in our proposed framework. Initially, raw data is embedded using a non-parametric UMAP approach. Additionally, UMAP is utilized to fit feature maps generated by ResNet50 models. These two plots together illustrate the transformation before and after learning, enhancing the interpretability of the learning process.

## 4. Results

This section contains the results of each stage of the proposed framework. All experiments are done in a PyTorch environment using Google Colab services and lo-

cal sources such as NVIDIA GeForce RTX 3050 Laptop GPU with 11.7 GB memory.

### *4.1. Stage 1. Low Resolution*

#### *4.1.1. Low Resolution Quality Control (LRQC)*

Experiments in low-resolution quality control (LRQC) mainly distinguished between good and bad labels using our VGG-like custom CNN. Consequently, two experiments with a different grouping of class labels were done. First, we test CNN performance in binary classification: Good and the other three classes. Second, 4-class classification is accomplished using all labels shown in Figure 3. Left-to-right flipping augmentation was applied for both experiments, and the same hyperparameters were used. These are listed in Table 2.

| Hyperparameters | Value |
|---|---|
| Optimizer | Adam |
| Learning Rate | 5 x 10-6 |
| Batch Size | 32 |
| Image Size | 128x128 px |
| Loss | Cross-Entropy |
| Loss Reduction | Mean |
| Number of Epochs | 10 |

Table 2: The hyperparameter/settings of LRQC.

*LRQC Binary Classification.* The subtask of direct binary classification between good and bad-quality images was successful. A validation set of 25 % was held out in training. Figure 14 shows results reaching 96 % validation accuracy in 10 epochs training next to the confusion matrix figure. These results demonstrated high performance in LQRC binary classification, encouraging further exploration into 4-class classification.

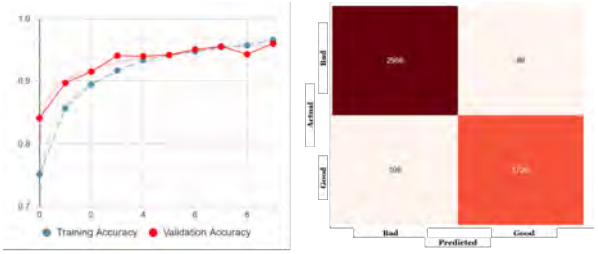



Figure 14: Results of binary LRQC. (left) Accuracy performance graph, (right) Confusion matrix of the model regarding four classes (right). 96 % validation accuracy was reached after 10 epochs of training. The model successfully distinguished classes from each other, promoting use in application.

*LRQC 4-Class Classification.* In this experiment, we ensure CNN's performance in group-based-4-fold cross-validation with randomized stratified split fashion. The groups indicate each eye folder to control possible data leakage, justified in Subsection 3.2.2. Figure 15 shows the main LRQC task results whereas Table 3 summarizes quantitative performance. The confusion matrix reveals that errors primarily occur between the No Cornea and No Liquid classes, indicating confusion between these classes by the model. Additionally, the recall scores for these classes are lower than those for Bad Coupling and Good. These findings suggest that improving LRQC classification specialization may be achievable by increasing the amount of data in minority classes.

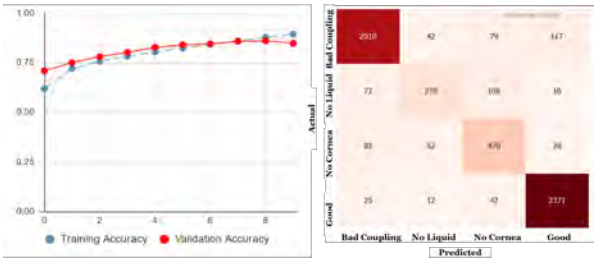



Figure 15: Results of 4-class LRQC. (left) Accuracy performance graph, (right) Confusion matrix of the model regarding four classes (right). 85 % validation accuracy was reached after 10 epochs of training. The model successfully distinguished Good from the other three labels, promoting use in application.

#### *4.1.2. Low Resolution Segmentation*

Segmentation was achieved for two different regions, the cornea and AC liquid of the eye, using two separate U-Net models. Left-to-right flipping augmentation was applied to increase dataset size. Hyperparameters are listed in Table 4. Figure 16 shows training and validation of Dice performance throughout the training. Figure 17 illustrates the resulting masks for the interested areas. These outcomes demonstrate the efficacy in finding the correct region of interest in AC using U-Net.

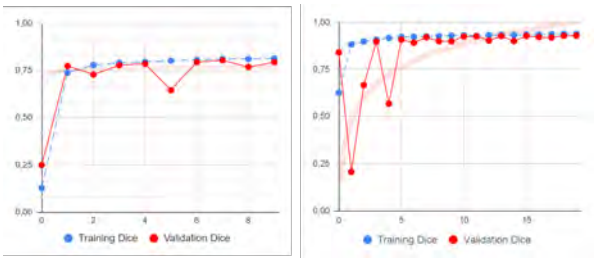



Figure 16: Results of LRS stage. 0.79 Dice score was reached after 10 epochs of training on the cornea (left). 0.93 Dice score was reached after 20 epochs of training on liquid (right).

### *4.2. Stage 2. Uveitis Screening*

The second stage is composed of two steps: (i) image level and (ii) eye level classifications from HR images of each eye.

| Class | **Specificity** | **Sensitivity (Recall)** | **Accuracy** |
|---|---|---|---|
| Bad Coupling | 0.95 ± 0.05 | 0.89 ± 0.04 | 0.93 ± 0.02 |
| No Liquid | 0.98 ± 0.01 | 0.58 ± 0.13 | 0.95 ± 0.02 |
| No Cornea | 0.96 ± 0.02 | 0.74 ± 0.2 | 0.93 ± 0.02 |
| Good | 0.95 ± 0.01 | 0.97 ± 0.02 | 0.96 ± 0.01 |

Table 3: Qualitative performance results of LRQC 4-class classification.

| **Hyperparameters** | **Value** |
|---|---|
| Optimizer | Adam |
| Learning Rate | 5 x 10-3 |
| Batch Size | 32 |
| Image Size | 128x128 px |
| Loss | Mean-Squared Error |
| Loss Reduction | Mean |

Table 4: The hyperparameter/settings of LRS for Cornea and Liquid segmentation models.

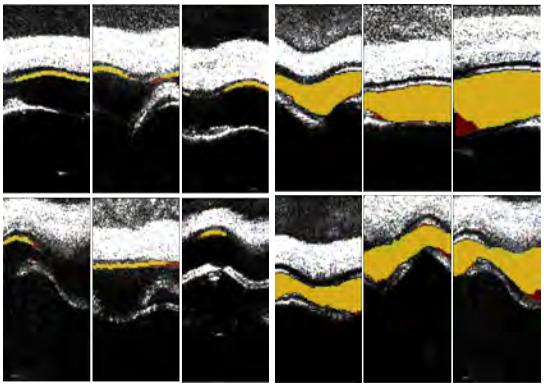


Figure 17: Visual results of Cornea (left) and AC Liquid (right) segmentation. (red) Ground truth, (orange) Prediction.

### *4.2.1. High-Resolution Binary Classification*

In HR Binary Classification, a Resnet50 pretrained on ImageNet was downloaded from the PyTorch server and fine-tuned with the setting described in Table 5. Left-to-right flipping augmentation was also applied to increase data size. Note that since we have relatively limited data and aim to build an end-to-end application, here the Leave-One-Eye-Out strategy is used to validate performance. This is further justified in Subsection 3.2.2.

Figure 18 shows confusion matrix result depicting the success in classifying Cells images from No Cells images despite the high false negative. The source of these false negatives was mainly SUN 0.5 graded eyes, which will be addressed in the next step. Moreover, Table 6 demonstrates validated performance scores for the fine-tuned models. Our fine-tuned Resnet50 models achieved high performance with 90.83 % accuracy and 97.06 % specificity. Figure 19 draws accuracy per eye-folder, counting how many frames/images matched correctly. None of the eye-folder had less than 50 %

| **Hyperparameters** | **Value** |
|---|---|
| Optimizer | SGD |
| Learning Rate | 5 x 10-4 |
| Batch Size | 16 |
| Weight Decay | 0.01 |
| SGD Momentum | 0.3 |
| Image Size | 224x224 px |
| Loss | Cross Entropy |
| Loss Reduction | Mean |
| Number of Epochs | 40 |

Table 5: The hyperparameter/settings of HR Binary Classification.

accuracy, showing all 26 folds were better than random guesses. Finally, Figure 20 shows a histogram of the probability distribution of the proposed binary classification model. This showcases the region/class in which the model's confidence accumulates. A potential threshold for class probability was found to be 0.5, as it seems to separate the two classes well.

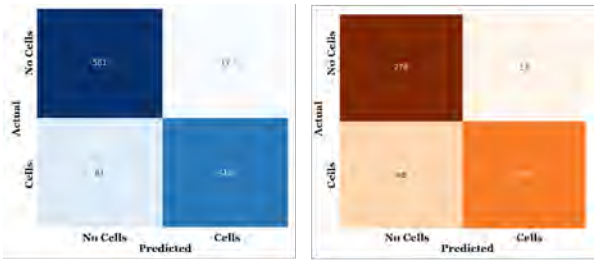



Figure 18: Confusion matrices of the High-Resolution Binary Classification of Cells and No Cells regarding (left) all eyes and (right) only SUN 0.5 graded eyes. 84 % of false negatives come from SUN 0.5 graded eyes.

### *4.2.2. Eye Level Diagnosis*

In the final diagnostic step, we experimented with two different voting schemes.

*Soft Voting.* Figure 21 shows the probability distribution map at eye level, i.e., average probabilities of eye folders. Note that colors represent their diagnostic (SUN-based) ground truth. Figure 22 shows the performance across different threshold operations and the confusion matrix at the ideal threshold. 0.2 was found to be the optimal threshold in this scheme, providing maximum recall. This indicates that for each eye to be

| | Specificity | Sensitivity (Recall) | F1-Score | Accuracy |
|---|---|---|---|---|
| All Eyes | 97.06 | 83.50 | 89.32 | 90.83 |
| **Only** SUN 0.5 graded eyes | 95.5 | 70.4 | 75.70 | 85.5 |

Table 6: Qualitative performance results of HR Binary classification.

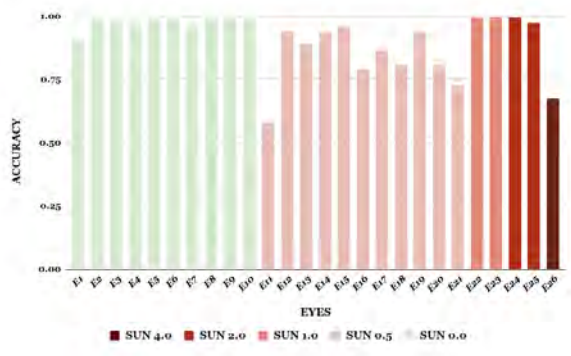



Figure 19: Accuracy per eye fold. No eyes' image-level accuracy is less than 50 %.

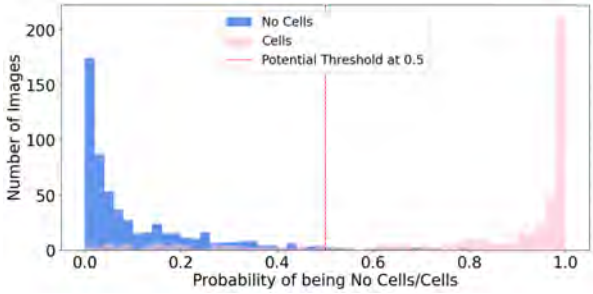



Figure 20: Probability distribution map as output of High-Resolution Binary Classification task at the image level.

diagnosed as Uveitis (SUN > 0), the average probability that images belonging to that eye contain WBC should be higher than 20 %. This voting mechanism resulted in 4 false negatives.

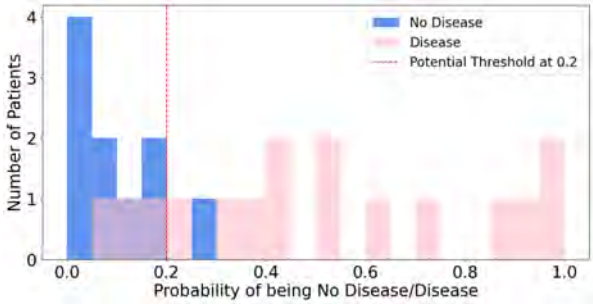



Figure 21: Probability distribution map as the output of Soft-Voting at the eye level.

*Hard Voting.* Figure 23 shows the performance across different threshold operations and the confusion matrix at the ideal threshold. 0.05 was found to be the optimal

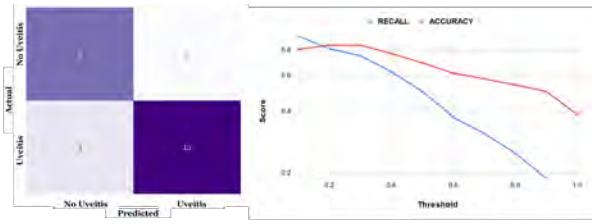



Figure 22: Quantitative results of soft voting scheme: (left) Confusion matrix and (right) Recall and accuracy performance versus cut-off threshold chosen.

threshold in this scheme, providing 100 % recall. This indicates that for each eye to be diagnosed as Uveitis (SUN > 0), 5 % of images belonging to that eye should be detected as Cells. This voting mechanism resulted in only 1 false positive, demonstrating the highest recall at Stage 2.

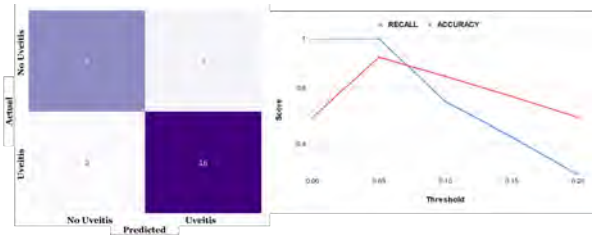



Figure 23: Quantitative results of the hard voting scheme: (left) Confusion matrix and (right) Recall and accuracy performance versus cut-off threshold (or percentage) chosen.

### *4.3. Stage 3. Explainable AI (xAI)*

The third and final stage utilizes GradCAM and UMAP methodologies to enhance the explainability and interpretability of the model.

*GradCAM.* The pure GradCAM technique (Selvaraju et al., 2020) was applied to Resnet50 models utilizing a Python library, specialized in explaining PyTorch-framework models (Gildenblat and contributors, 2021). A total of four cases are reported in this paper to showcase findings from GradCAM.

Figure 24 encapsulates GradCAM application on positive class, i.e., the class where WBC presence should be found. We include success and failure cases, one from each. The first case depicts that Cells are found in AC liquid with high confidence, supporting our hypothesis that WBC morphology appears as trace-like structures in the ultrasound. Contrarily, the failed case

shows it is still difficult to distinguish WBC morphology from the cornea or other eye structures that can be captured in the same frame.

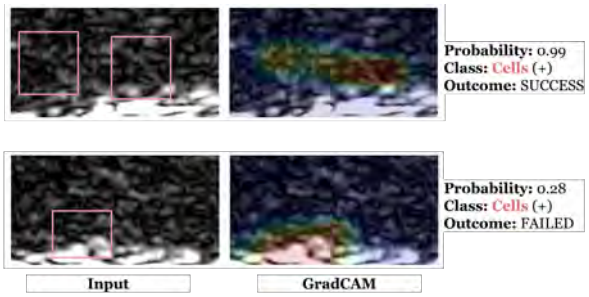



Figure 24: White blood cells were seen in an eye with cells with SUN 0.5 (accuracy of 58 %). The lens can lower the model's confidence in seeing cells.

Figure 25 illustrates the finding on the No Cells group in the same fashion. The case in which the model predicted a false negative shows that artifacts, probably caused by movement, can be mistaken as WBC presence.

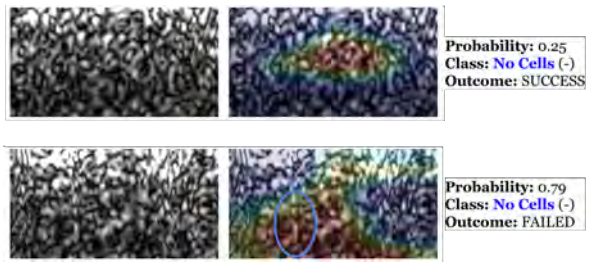



Figure 25: Anterior chamber visualization of an eye with no cells with SUN 0.0 (accuracy of 92%). Artifacts due to motion during examination can mislead the model.

*UMAP.* A UMAP was fit on (i) raw data and (ii) Resnet50 feature maps. The number of neighbors was set to 5. The minimum distance was kept at default, 0.1. These settings define how widespread the reduction will be.

Figure 26 shows the first plot colored by both WBC presence ground truth and SUN-grades of their belonging eye. Despite the control (No Cells) group being approximately clustered, the positive (Cells) group has a high outlier percentage. This validates the concern about achieving high recall, i.e., a low number of false negatives. This is further supported by the SUN grade-colored reduction plot, where SUN 0.5 graded eyes are mashed with the SUN 0 graded eyes. Furthermore, the impact of learning is shown in Figure 27, where Cells and No Cells classes are separated better. Note that the SUN grade-colored plot shows the difficulty of distinguishing SUN 0.5 from SUN 0.0, even after learning the presence of cells. This validates our need for an eye-level diagnosis scheme.

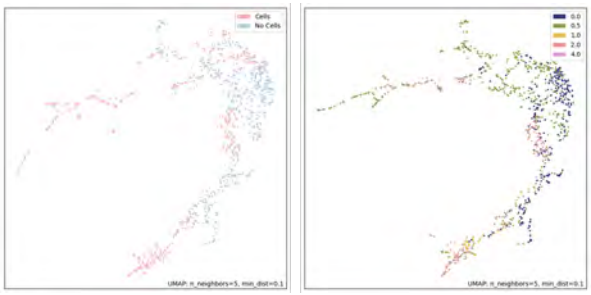


Figure 26: UMAP embedding on flattened image data colored by WBC presence (left) and SUN grade (right).

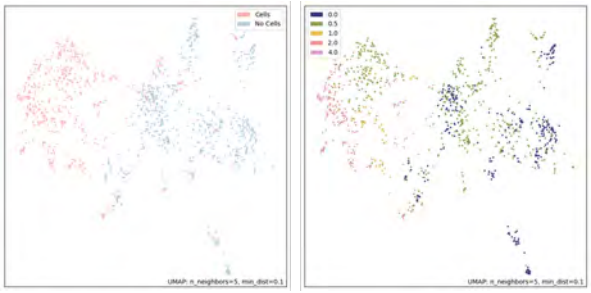


Figure 27: UMAP embedding on Resnet50 feature maps colored by WBC presence (left) and SUN grade (right).

## 5. Discussion

This study serves as a milestone in building an intelligent, explainable, and accessible end-to-end uveitis screening utilizing a novel ultrasound imaging technology. In this discussion, we will critically analyze the implications of our results, consider the limitations of the proposed technology & framework, and suggest directions for future work. By examining these aspects, we aim to offer a balanced perspective that underscores the strengths and constraints of our proposed framework. Embarking on this effort ensures that 2D Ultrasound technology becomes a cornerstone in enhancing uveitis diagnostics capabilities within clinics. Thereby it amplifies accessibility to automated eye care, particularly in the developing world, and for individuals in vulnerable demographics such as children and those from economically disadvantaged backgrounds. Democratization of access to sufficient uveitis screening with a Neosonics® device employing a deep learning-based framework also unlocks the ethical and critical AI application in healthcare.

Our framework has shown promising results in every stage of the study. In Stage 1 Low Resolution, image level accuracy for LRQC has reached 85 %. Identifying regions of the cornea and liquid in the eye was also accomplished with high Dice scores of 0.79 and 0.93, respectively. In Stage 2 Uveitis Screening, where we deal with HR images, first a binary classification for WBC presence at the image level was achieved with 90.83 % accuracy and 89.32 % F1 score. Second, hard voting-

based diagnosis was set up with customized thresholds. This resulted in 25/26 accuracy at the eye level, demonstrating the proposed framework's efficacy *in vivo*. In the final stage, the HR binary classification model was subject to xAI. First, GradCAM was deployed to unpack features learned by the deep learning model. Patterns to identify WBC were visualized by GradCAM heatmaps. Furthermore, these heatmaps could point out what was 'confusing' for the model as WBC presence. Second, a dimensionality reduction technique, namely UMAP, was applied to data and extracted features by deep learning, showing that the algorithm can efficiently separate binary classes.

### *5.1. Limitations*

The main limitations of the proposed framework are listed in this subsection.

1. *Noise and artifacts* Ultrasound is prone to noise and clutter. Typically, this positions it more as a tool for initial screening rather than a method for definitive diagnosis. Due to the novelty of the Neosonics® US device, its unique noise-related problems are yet to be addressed.
2. *Imaging protocol*. Another constraint is to generate data in a standardized way. This further gets complicated as the device is developing and many scans are being conducted to develop the ideal protocol.
3. *Ground truth generation*. During the conduction of the study, annotation-based ground truths for the low-resolution stage were done manually and are prone to human error. Moreover, only one medical expert examined the eyes to provide a clinical diagnosis. This further risks the possibility of specializing in uveitis grading.
4. *Data Scarcity.* The HR stage was tested on 26 eyes from 20 participants who visited the same hospital. Their demographics were not disclosed in this report and the generalizability of this work remains unexplored due to the limited dataset.

### *5.2. Future Work*

The future steps of this research involve first increasing the dataset size by acquiring *in vivo* data or including *in vitro* data techniques in the framework. Second, eye-level diagnosis can be enriched by introducing ensemble techniques to specialize cell counting, grading eyes between SUN 0.5 - 4.

Moreover, attention-based training mechanisms can be useful. These mechanisms would further utilize GradCAM heatmaps to learn efficiently. Data enhancement techniques such as denoising, morphological preprocessing, and blur filtering could be experimented with to see their impacts on each deep learning task separately.

In summary, the results obtained throughout the various stages of our study demonstrate the robustness and efficacy of our proposed framework in advancing uveitis screening technology. The high accuracy rates achieved in image-level classification, coupled with the successful identification of key regions within the eye, underscore the potential of our approach to enhance diagnostic accuracy and efficiency. Furthermore, the deployment of advanced techniques such as GradCAM and UMAP has provided valuable insights into the inner workings of our deep learning model and its learning mechanisms. These findings not only contribute to the field of uveitis diagnostics but also pave the way for the ethical and effective integration of AI in healthcare. Moving forward, further validation and refinement of our framework hold promise for revolutionizing uveitis screening and improving patient outcomes on a global scale.

## Acknowledgments

This study rigorously followed ethical guidelines, ensuring patient confidentiality and anonymization of data. Data was acquired *in vivo* at *Hospital Germans Trias i Pujol* in Barcelona, Spain. Medical consultations of the Ophthalmology unit were insightful and invaluable. KRIBA.AI copyrights all technological advances, however, the authors of this thesis sought no profit from this research. This study was carried out with the assistance of Francesc Carandell Verdaguer and Beatrice Jobst, who are data scientists at KRIBA.AI. We are grateful for their personal and professional efforts and want to acknowledge their contributions.

## References

Abd El Latif, E., Fayez Goubran, W., El Gemai, E.E.D.M., Habib, A.E., Abdelbaki, A.M., Ammar, H., Seleet, M., 2019. Pattern of childhood uveitis in egypt. Ocular Immunology and Inflammation 27, 883–889. doi:10.1080/09273948.2018.1502325.

Ajanovic, S., Jobst, B., Jimenez, J., Quesada, R., Santos, F., Lopez-Azorín, M., Valverde, E., Ybarra, M., Bravo, M., Petrone, P., Sial, H., Muñoz, D., Agut, T., Salas, B., Carreras, N., Alarcon, A., Iriondo, M., Luaces, C., Ibañez, A., Bassat, Q., 2023. Meningitis screening in young infants based on a novel, non-invasive, transfontanellar ultrasound device: a proof-of-concept study doi:10.21203/rs.3.rs-3677475/v1. in review, 2024.

Al-Ani, H.H., Sims, J.L., Tomkins-Netzer, O., Lightman, S., Niederer, R.L., 2020. Vision loss in anterior uveitis. The British journal of ophthalmology 104, 1652–1657. doi:10.1136/bjophthalmol-2019-315551.

Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I., on behalf of the Precise4Q consortium, 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 20, 310. doi:10.1186/s12911-020-01332-6.

Anderson, B.O., Braun, S., Carlson, R.W., Gralow, J.R., Lagios, M.D., Lehman, C., Schwartsmann, G., Vargas, H.I., 2003. Overview of breast health care guidelines for countries with limited resources. The breast journal 9, S42–S50. doi:10.1046/j.1524-4741.9.s2.3.x.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion 58, 82–115. URL: https://www.sciencedirect.com/science/article/pii/S1566253519308103, doi:10.1016/j.inffus.2019.12.012.

Bélard, S., Tamarozzi, F., Bustinduy, A.L., Wallrauch, C., Grobusch, M.P., Kuhn, W., Brunetti, E., Joekes, E., Heller, T., 2016. Point-of-care ultrasound assessment of tropical infectious diseases–a review of applications and perspectives. The American journal of tropical medicine and hygiene 94, 8–21. doi:10.4269/ajtmh.15-0421.

Bodaghi, B., Cassoux, N., Wechsler, B., Hannouche, D., Fardeau, C., Papo, T., Huong, D.L., Piette, J.C., LeHoang, P., 2001. Chronic severe uveitis: etiology and visual outcome in 927 patients from a single center. Medicine 80, 263–270. doi:10.1097/00005792-200107000-00005.

Buda, N., Segura-Grau, E., Cylwik, J., Wełnicki, M., 2020. Lung ultrasound in the diagnosis of covid-19 infection - a case series and review of the literature. Advances in medical sciences 65, 378–385. doi:10.1016/j.advms.2020.06.005.

Chang, J., Mccluskey, P., Wakefield, D., Pleyer, U., Forrester, J., 2008. Acute Anterior Uveitis and HLA-B27: What's New? Springer. doi:10.1007/978-3-540-69459-5_2.

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. doi:10.1007/s10278-017-9997-y.

Chou, T.H., Yeh, H.J., Chang, Chun-Chaoa, b., Tang, J.H., Kao, Wei-Yua, b., Su, I.C., Li, C.H., Chang, W.H., Huang, C.K., Sufriyana, Herdiantrid, e., Su, Emily Chia-Yud, f., 2021. Deep learning for abdominal ultrasound: A computer-aided diagnostic system for the severity of fatty liver. Journal of the Chinese Medical Association 84, 842–850. doi:10.1097/JCMA.0000000000000585.

Cinà, G., Röber, T., Goedhart, R., Birbil, I., 2022. Why we do need explainable ai for healthcare. doi:10.48550/arXiv.2206.15363.

Dandona, L., Dandona, R., John, R.K., McCarty, C.A., Rao, G.N., 2000. Population based assessment of uveitis in an urban population in southern india. The British journal of ophthalmology 84, 706–709. doi:10.1136/bjo.84.7.706.

Darrell, R.W., Wagener, H.P., Kurland, L.T., 1962. Epidemiology of uveitis. incidence and prevalence in a small urban community. Archives of ophthalmology (Chicago, Ill. : 1960) 68, 502–514. doi:10.1001/archopht.1962.00960030506014.

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, Miami, FL, USA. pp. 248–255. doi:10.1109/CVPR.2009.5206848.

Deuchler, S., Dail, Y.A., Koch, F., Buedel, C., Ackermann, H., Flockerzi, E., Seitz, B., 2023. Efficacy of simulator-based slit lamp training for medical students: A prospective, randomized trial. Ophthalmology and Therapy 12, 2171–2186. doi:10.1007/s40123-023-00733-w.

Dietrich, C., Sirlin, C., O'Boyle, M., Dong, Y., Jenssen, C., 2019. Editorial on the current role of ultrasound. Applied Sciences 9, 3512. doi:10.3390/app9173512.

ten Doesschate, J., 1982. Causes of blindness in the netherlands. Documenta ophthalmologica. Advances in ophthalmology 52, 279–285. doi:10.1007/BF01675857.

Edelsten, C., Reddy, M.A., Stanford, M.R., Graham, E.M., 2003. Visual loss associated with pediatric uveitis in english primary and referral centers. American journal of ophthalmology 135, 676–680. doi:10.1016/s0002-9394(02)02148-7.

Elvira, L., Fernandez, A., León, L., Ibáñez, A., Parrilla, M., Martinez-Graullara, O., Jimenez, J., 2023. Evaluation of the cell concentration in suspensions of human leukocytes by ultrasound imaging: The influence of size dispersion and cell type. Sensors 23, 977. doi:10.3390/s23040977.

Eser-Ozturk, H., Sullu, Y., 2020. Pediatric uveitis in a referral center in north part of turkey. Ocular Immunology and Inflammation 29, 1299–1303. doi:10.1080/09273948.2020.1758158.

Fledelius, H.C., 1996. Ultrasound in ophthalmology. Ultrasound in Medicine & Biology 23, 365–375. doi:10.1016/S0301-5629(96)00213-X.

Gildenblat, J., contributors, 2021. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam.

Häring, G., et al., 1998. Ultrasound biomicroscopic imaging in intermediate uveitis. The British Journal of Ophthalmology 82, 625–629. doi:10.1136/bjo.82.6.625.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. doi:10.48550/arXiv.1512.03385.

Hoffmann, B., Schafer, J.M., Dietrich, C.F., 2020. Emergency ocular ultrasound - common traumatic and non-traumatic emergencies diagnosed with bedside ultrasound. Ultraschall in der Medizin (Stuttgart, Germany : 1980) 41, 618–645. doi:10.1055/a-1246-5984.

Jabs, D.A., Dick, A., Doucette, J.T., Gupta, A., Lightman, S., McCluskey, P., Okada, A.A., Palestine, A.G., Rosenbaum, J.T., Saleem, S.M., Thorne, J., Trusko, B., of Uveitis Nomenclature Working Group, S., 2018. Interobserver agreement among uveitis experts on uveitic diagnoses: The standardization of uveitis nomenclature experience. American journal of ophthalmology 186, 19–24. doi:10.1016/j.ajo.2017.10.028.

Jennings, C.M., King, J.B., Parekh, S.H., 2022. Low-cost, minimalistic line-scanning confocal microscopy. Opt. Lett. 47, 4191–4194. doi:10.1364/OL.456347.

Jimenez, X., Shukla, S.K., Ortega, I., Illana, F.J., Castro-González, C., Marti-Fuster, B., Butterworth, I., Arroyo, M., Anthony, B., Elvira, L., 2016. Quantification of very low concentrations of leukocyte suspensions in vitro by high-frequency ultrasound. Ultrasound in Medicine & Biology 42, 1568–1573. doi:10.1016/j.ultrasmedbio.2016.01.027.

Kaur, K., Gurnani, B., 2024. Slit-lamp biomicroscope. Updated 2023 Jun 11. In: Stat Pearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK587440/.

Kempen, J.H., Ganesh, S.K., Sangwan, V.S., Rathinam, S.R., 2008. Interobserver agreement in grading activity and site of inflammation in eyes of patients with uveitis. American journal of ophthalmology 146, 813–8.e1. doi:10.1016/j.ajo.2008.06.004.

Konstantopoulou, K., Del'omo, R., Morley, A., Karagiannis, D., Bunce, C., Pavesio, C., 2012. A comparative study between clinical grading of anterior chamber flare and flare reading using the kowa laser flare meter. International Ophthalmology doi:10.1007/s10792-012-9616-3.

Krichen, M., 2023. Convolutional neural networks: A survey. Computers 12, 151. doi:10.3390/computers12080151.

Krumpaszky, H.G., Klauss, V., 1992. Erblindungsursachen in bayern. auswertung einer repräsentativen stichprobe der blindengeldakten aus oberbayern. Klinische Monatsblatter fur Augenheilkunde 200, 142–146. doi:10.1055/s-2008-1045729.

Lee, J.G., Jun, S., Cho, Y.W., Lee, H., Kim, G.B., Seo, J.B., Kim, N., 2017. Deep learning in medical imaging: General overview. Korean Journal of Radiology 18, 570–584. doi:10.3348/kjr.2017.18.4.570.

Lee, J.H., Boning, D.S., Anthony, B.W., 2018. Measuring the absolute concentration of microparticles in suspension using high-frequency b-mode ultrasound imaging. Ultrasound in Medicine & Biology 44, 1086–1099. doi:10.1016/j.ultrasmedbio.2018.01.005.

Lin, B.S., Chen, J.L., Tu, Y.H., Shih, Y.X., Lin, Y.C., Chi, W.L., Wu, Y.C., 2020. Using deep learning in ultrasound imaging of bicipital peritendinous effusion to grade inflammation severity. IEEE journal of biomedical and health informatics 24, 1037–1045. doi:10.1109/JBHI.2020.2968815.

Liu, X., Song, J.L., Wang, S.H., Zhao, J.W., Chen, Y.Q., 2017. Learning to diagnose cirrhosis with liver capsule guided ultrasound image classification. Sensors (Basel, Switzerland) 17, 149. doi:10.3390/s17010149.

Liu, X., et al., 2020. Instrument-based tests for measuring anterior chamber cells in uveitis: A systematic review. Ocular Immunology and Inflammation 28, 898–907. doi:10.1080/09273948.2019.1640883.

Maring, M., Saraf, S.S., Blazes, M., Sharma, S., Srivastava, S., Pepple, K.L., Lee, C.S., 2022. Grading anterior chamber inflammation with anterior segment optical coherence tomography: An overview. Ocular immunology and inflammation 30, 357–363. doi:10.1080/09273948.2022.2028289.

McInnes, L., Healy, J., Melville, J., 2018. Umap: Uniform manifold approximation and projection. Journal of Open Source Software 3, 861. doi:10.21105/joss.00861.

Meng, D., Zhang, L., Cao, G., Cao, W., Zhang, G., Hu, B., 2017. Liver fibrosis classification based on transfer learning and fcnet for ultrasound images. IEEE Access 5, 5804–5810. doi:10.1109/ACCESS.2017.2689058.

Miserocchi, E., Fogliato, G., Modorati, G., Bandello, F., 2013. Review on the worldwide epidemiology of uveitis. European Journal of Ophthalmology 23, 705–717. doi:10.5301/ejo.5000278.

Okonkwo, O., Hassan, A., Bogunjoko, T., Akinye, A., Akanbi, T., Agweye, C., 2023. Low rates of optical coherence tomography utilization in the diagnosis and management of retinovascular diseases in a lower middle-income economy. Nigerian Journal of Clinical Practice 26, 1011–1016. doi:10.4103/njcp.njcp_911_22.

Orlando, N., Gillies, D.J., Gyacskov, I., Romagnoli, C., D'Souza, D., Fenster, A., 2020. Automatic prostate segmentation using deep learning on clinically diverse 3d transrectal ultrasound images. Medical Physics 47, 2413–2426. doi:10.1002/mp.14134.

Ortiz-González, L., Ortiz-Peces, C., Calle-Guisado, V., Ortiz-Peces, L., 2024. Ultrasound diagnosis of anterior uveitis in primary care. Anales de Pediatría 100, 380–381. doi:10.1016/j.anpede.2024.04.003.

Päivönsalo-Hietanen, T., Tuominen, J., Saari, K.M., 2000. Uveitis in children: population-based study in finland. Acta ophthalmologica Scandinavica 78, 84–88. doi:10.1034/j.1600-0420.2000.078001084.x.

de Parisot, A., Jamilloux, Y., Kodjikian, L., Errera, M.H., Sedira, N., Heron, E., Pérard, L., Cornut, P.L., Schneider, C., et al., 2020. Evaluating the cost-consequence of a standardized strategy for the etiological diagnosis of uveitis (ulisse study). PLoS One 15, e0228918. doi:10.1371/journal.pone.0228918.

Pistilli, M., Gangaputra, S.S., Pujari, S.S., Jabs, D.A., Levy-Clarke, G.A., Nussenblatt, R.B., Rosenbaum, J.T., Sen, H.N., Suhler, E.B., Thorne, J.E., et al., 2021. Contemporaneous risk factors for visual acuity in non-infectious uveitis. Ocular immunology and inflammation 29, 1056–1063. URL: https://doi.org/10.1080/09273948.2020.1828493, doi:10.1080/09273948.2020.1828493.

Qian, R., McNabb, R.P., Zhou, K.C., Mousa, H.M., Saban, D.R., Perez, V.L., Kuo, A.N., Izatt, J.A., 2021. In vivo quantitative analysis of anterior chamber white blood cell mixture composition using spectroscopic optical coherence tomography. Biomedical optics express 12, 2134–2148. doi:10.1364/BOE.419063.

Raheem, A., 2021. Effects of artifacts on the diagnosis of ultrasound image. Medico Legal Update 21, 327–336. doi:10.37506/mlu.v21i4.3152.

Rix, A., Lederle, W., Theek, B., Lammers, T., Moonen, C., Schmitz, G., Kiessling, F., 2018. Advanced ultrasound technologies for diagnosis and therapy. Journal of Nuclear Medicine 59, 740–746. doi:10.2967/jnumed.117.200030.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. doi:10.48550/arXiv.1505.04597.

Rothova, A., Suttorp-van Schulten, M.S., Frits Treffers, W., Kijlstra, A., 1996. Causes and frequency of blindness in patients with intraocular inflammatory disease. The British journal of ophthalmology 80, 332–336. doi:10.1136/bjo.80.4.332.

Sainburg, T., McInnes, L., Gentner, T.Q., 2021. Parametric umap embeddings for representation and semisupervised learning. Neural Computation 33, 2881–2907. doi:10.1162/neco_a_01434.

Seepongphun, U., Sittivarakul, W., Dangboon, W., Chotipanvithayakul, R., 2021. The pattern of uveitis in a pediatric population at a tertiary center in thailand. Ocular Immunology and Inflammation 31, 56–64. doi:10.1080/09273948.2021.1980814.

Selvaraju, R., Cogswell, M., Das, A., et al., 2020. 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.

Shi, G., Jiang, Z., Deng, G., Liu, G., Zong, Y., Jiang, C., Chen, Q., Lu, Y., Sun, X., 2019. Automatic classification of anterior chamber angle using ultrasound biomicroscopy and deep learning. Trans. Vis. Sci. Tech. 8, 25. doi:10.1167/tvst.8.4.25.

Sial, H., Carandell, F., Ajanovic, S., Jiménez, J., Quesada, R., Santos, F., Buck, W.C., UNITED Study Consortium, Bassat, Q., Jobst, B., Petrone, P., 2024. Deep learning framework for non-invasive infant meningitis screening from high-resolution ultrasound images In review, 2024.

Srinivasu, P.N., Sandhya, N., Jhaveri, R., Raut, R., 2022. From blackbox to explainable ai in healthcare: Existing tools and case studies. Mobile Information Systems , 1–20doi:10.1155/2022/8167821.

St. Croix, C.M., Shand, S.H., Watkins, S.C., 2005. Confocal microscopy: Comparisons, applications, and problems. BioTechniques 39, S2–S5. doi:10.2144/000112089.

Suttorp-Schulten, M., Rothova, A., 1996. The possible impact of uveitis in blindness: a literature survey. The British journal of ophthalmology 80, 844–848. doi:10.1136/bjo.80.9.844.

Tabbut, M., Bates, A., Marple, G., Gramer, D., Tabbut, B., 2019. Point-of-care ultrasound in the evaluation of the acutely painful red eye. The Journal of Emergency Medicine 57, 705–709. doi:10.1016/j.jemermed.2019.04.034.

Trusko, B., Thorne, J., Jabs, D., Belfort, R., Dick, A., Gangaputra, S., Nussenblatt, R., Okada, A., Rosenbaum, J., Standardization of Uveitis Nomenclature (SUN) Project, 2013. The standardization of uveitis nomenclature (sun) project. development of a clinical evidence base utilizing informatics tools and techniques. Methods of information in medicine 52, 259–S6. doi:10.3414/ME12-01-0063.

Uçar, E., 2022. Classification of myositis from muscle ultrasound images using deep learning. Biomedical Signal Processing and Control 71, 103277. doi:10.1016/j.bspc.2021.103277.

Wang, W., Wang, L., Wang, X., Zhou, S., Yang, J., 2021a. A deep learning system for automatic assessment of anterior chamber angle in ultrasound biomicroscopy images. Trans. Vis. Sci. Tech. 10, 21. doi:10.1167/tvst.10.11.21.

Wang, Y., Tang, C., Wang, J., Sang, Y., Lv, J., 2021b. Cataract detection based on ocular b-ultrasound images by collaborative monitoring deep learning. Knowledge-Based Systems 231, 107442. doi:10.1016/j.knosys.2021.107442.

Wu, W.T., Chang, K.V., Hsu, Y.C., Hsu, P.C., Ricci, V., Özçakar, L., 2020. Artifacts in musculoskeletal ultrasonography: From physics to clinics. Diagnostics (Basel, Switzerland) 10, 645. doi:10.3390/diagnostics10090645.

Yang, S., Lemke, C., Cox, B.F., Newton, I.P., Näthke, I., Cochran, S., 2021. A learning-based microultrasound system for the detection of inflammation of the gastrointestinal tract. IEEE Transactions on Medical Imaging 40, 38–47. doi:10.1109/TMI.2020.3021560.

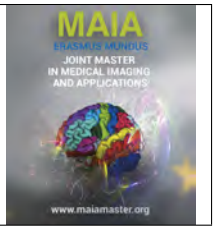
Yi, J., Kang, H.K., Kwon, J.H., Kim, K.S., Park, M.H., Seong, Y.K., Kim, D.W., Ahn, B., Ha, K., Lee, J., Hah, Z.H., Bang, W.C., 2021. Technology trends and applications of deep learning in ultrasonography: Image quality enhancement, diagnostic support, and improving workflow efficiency. Ultrasonography (Seoul, Korea) 40, 7–22. doi:10.14366/usg.20102.

Zhang, X., Lv, J., Zheng, H., Sang, Y., 2020. Attention-based multimodel ensemble for automatic cataract detection in b-scan eye ultrasound images, in: 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK. pp. 1–10. doi:10.1109/IJCNN48605.2020.9207696.

Zur, D., Neudorfer, M., Shulman, S., Rosenblatt, A., Habot-Wilner, Z., 2016. High-resolution ultrasound biomicroscopy as an adjunctive diagnostic tool for anterior scleral inflammatory disease. Acta Ophthalmologica 94, e384–e389. doi:10.1111/aos.12995.

Medical Imaging and Applications

Master Thesis, June 2024




# Automated Segmentation of White Matter Hyperintensities using Deep Learning

Edwing Ulin[a,b], Santiago Estrada[a,b], Martin Reuter[b,c,d]

*eulinbriseno@gmail.com, santiago.estrada@dzne.de, martin.reuter@dzne.de*

[a] *Population Health Sciences, German Center for Neurodegenerative Diseases (DZNE),Bonn, Germany*
[b] *AI in Medical Imaging, German Center for Neurodegenerative Diseases (DZNE),Bonn, Germany*
[c] *A.A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston MA, USA*
[d] *Deparment of Radiology, Harvard Medical School, Boston MA, USA*

**Abstract**

This thesis presents an in-depth analysis of automated segmentation techniques for White Matter Hyperintensities (WMH) in brain magnetic resonance images (MRIs), which are essential biomarkers in the study of neurodegenerative diseases such as Alzheimer's Disease. Utilizing advanced deep learning architectures, specifically Dense U-Net, FastSurferCNN, and nnU-Net, this work assesses the performance of these models on multimodal MRI data in two significant cohorts, the Rhineland Study and the UK Biobank. The rationale for evaluating different architectures lies in their unique design principles and capabilities. Dense U-Net utilizes densely connected layers to enhance feature learning, critical for the detailed and varied imaging characteristics of WMH. FastSurferCNN, designed for speed and efficiency, enables rapid processing that is vital for clinical application and large dataset analysis. nnU-Net, with its self-configuring networks, provides flexibility in handling diverse imaging conditions without extensive manual tuning, ensuring robustness across different data types. These diverse strengths necessitate a comparative analysis to identify the most effective model for accurate and efficient WMH segmentation. The research methodology employs a rigorous, comparative analysis of the three deep learning models using a variety of input configurations and kernel sizes to determine the most effective approaches for WMH detection and quantification. Transfer learning techniques are employed, leveraging pseudo-labeled data from a large independent cohort to enhance the models' learning efficacy and adaptability. This approach enables a detailed examination of the models' abilities to process diverse and intricate imaging datasets. The outcomes of this work suggest that all benchmarked methods can detect WMHs in scans that follow the distribution of the training data. However, all methods struggle with segmenting out-of-domain scans. These results indicate that there is still a need to further develop WMH detection methods that can generalize better and support a wider range of input scans.

*Keywords:* White Matter Hyperintensities, Deep Learning, MRI Segmentation, Dense U-Net, FastSurferCNN, nnU-Net, Neuroimaging, Alzheimer's Disease, Comparative Analysis, Multi-planar MRI Data, Automated Segmentation, Neurodegenerative Diseases, Image Processing

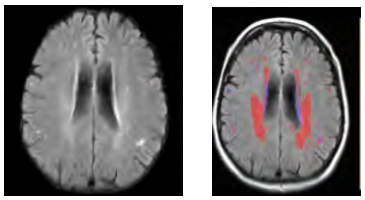
## 1 Introduction

### *1.1 Motivation*

The role of neuropathology in the study of Alzheimer's Disease mainly consisted of the analysis of postmortem diagnosis; however, advances in clinical diagnosis and pathophysiology redefine the importance of neuropathologic evaluation of the brain. At the same, a massive shift in the importance of the pathological mechanism underlying Alzheimer's Disease develops decades before the significant symptoms of the disease are evident (Trejo-Lopez et al., 2022). Hence, the early detection of lesions in the brain's structure influences the diagnosis of AD. According to recent research, neuropathological changes such as amyloid plaques and neurofibrillary tangles begin to form long before clinical symptoms appear (Scheltens et al., 1992). The white matter hyperintensities (WMH) of presumed vascular

origin, commonly refer as leukoaraiosis[1], are heavily linked to MRI of older subjects and patients with neurodegenerative diseases. These lesions consist of demyelination and axonal degeneration with high interaction with other pathological features, such as secondary cortical and long tract damage, and contribute to accumulating brain damage (Debette and Markus, 2010; Scheltens et al., 1992). Therefore, the early detection of this lesion provides a significant opportunity to prevent or reverse brain damage and mitigate cognitive deterioration (Scheltens et al., 1992). The lack of pathology studies compared with the number of WMH captured in imaging studies is related to the difficulty in matching up the individual small lesions on imaging with their pathological counterpart since the end stages of the disease obliterate the earliest stages.

Historically, WMH were associated primarily with demyelination and axonal loss. However, more recent studies utilizing MRI have demonstrated that WMH are also linked to microglial and endothelial activation. This suggests that periventricular and deep WMH are part of a continuous pathological process rather than distinct conditions (Scheltens et al., 1992). In the broadly used imaging method, MRI structural changes in the integrity of the brain's white matter. The images demonstrate the heterogeneity of the WMH as the amounts of damage it shows in Figure 1, easily recognized by the degree of "whiteness" related to the increased water content and mobility, demyelination, and axonal loss (Wardlaw et al., 2015).



(a) (b)

Figure 1: a) a normal Flair image with intense and low WMH damage b) damage segmentation in blue and red respectively

The relevance of this lesion, since it can be measured quantitatively in a non-invasive manner was proposed by Debette and Markus (2010) as an intermediated marker to identify a new risk factor for clinical trials. The high clinical importance of this lesion caused the development of computational methods for quantifying the WMH volume. This has led many research groups to develop their own scanning protocols and segmenting algorithms. Despite these advances, there remains a critical need for more comprehensive validation to ensure that these protocols and algorithms remain accurate, reliable, and generalized across contexts and applications. The manual measurement of the WMH volume requires a huge investment of time, hence the need for automated approaches (Scheltens et al., 1992; Wardlaw et al., 2015).

One significant limitation of contemporary segmentation techniques is their lack of accessibility via open-source software. This restricts the ability of researchers and practitioners to replicate and validate findings independently and may result in the propagation of inaccurate results. Furthermore, numerous available techniques require extensive tuning of their hyperparameters, a process that can be complex and time-consuming. Moreover, the open-source software must exhibit more generalizability across diverse datasets, impeding its practical utility and resilience in real-world settings. Therefore, it is critical to develop a validated tool that is open-source and capable of generalizing across various datasets, ensuring reliability and ease of use in various applications.

### *1.2 Related work*

Early methods for automated segmentation utilized features related to intensity and shape, such as k-nearest neighbor classification with tissue type priors (Steenwijk et al., 2013), *Lesion-TOpology-preserving Anatomical Segmentation* atlas-based (Shiee et al., 2014), the *Lesion Segmentation Prediction Algorithm* based on pixel density (Shiee et al., 2010), and the *Lesion Segmentation Tool Lesion Growth Algorithm* (Schmidt et al., 2012) a deformation field segmentation method. These techniques, achieved reasonable results as described in Heinen et al. (2019). However, these algorithms tend to operate slowly, necessitate building models from the ground up for each new dataset.

Conversely, Tran et al. (2022) conducted an analysis of various methodologies employed in the segmentation process. The first is an advanced version of the *White matter Hyperintensities Automated Segmentation Algorithm* (WHASA) (Tran et al., 2022), which combines non-linear diffusion and watershed segmentation to delineate regions suspected them to be merged based on intensity similarities, and candidate regions are identified using both intensity and spatial rules. Also, a k-nearest neighbors approach within the FSL framework, the *Brain Intensity AbNormality Classification Algorithm* (BIANCA) by Griffanti et al. (2016), classified the voxel based on the intensity and spatial features to produce a probability map of the lesion presence. Finally, *nicMLesion* (Valverde et al., 2019) is a two-cascading convolutional neural network (CNN) designed to be sensitive to lesions and specific to false positives (Tran et al., 2022).

In contrast, Park et al. (2021) presented the use of a variant a U-Net with the inclusion of the multi-scale highlighting foregrounds in a 2D network with data augmentation and ensemble consisting of 5-fold cross-validation resulting in 5 models during training with a

[1] a particular abnormal change in the appearance of white matter near the lateral ventricles

majority voting output. Furthermore, Isensee et al. (2021) developed a 2D generic UNet, commonly used as a benchmark in medical segmentation. This benchmark method relies on the easy adaptation of the Net to every new dataset, using domain knowledge such as the fixed design of non-dataset related dataset inputs and the use of a selection of parameters to optimize the Net from the *dataset fingerprint*. At last, Henschel et al. (2020) presented a fully automated pipeline for neuroimaging. This open-source project used the advantage of dense modules as feature extractors and the CNN configuration for a faster training process. The contributions of this work are the use of max out instead of stacking the outputs of previous layers and the use of 3 pipelines for the analysis in the different planes. The SHIVA method from the Early detection of white matter hyperintensities using SHIVA-WMH detector by Tsuchida et al. (2023) featured a 3D U-Net architecture with the inclusion of the dropout rates in order to optimize the performance.

The Medical Image Computing and Computer-Assisted Intervention Society (MICCAI) is at the forefront of promoting innovation in medical imaging through its annual challenges. The specific inclusion of tasks for WMH segmentation in recent challenges highlights the critical need for advancements in this area. The MICCAI 2017 challenge on Brain Lesions (BrainLes) focused on various brain lesions. It emphasized white matter hyperintensities, driving the creation of algorithms that improve the accuracy and reliability of segmentation methods. Critical contributions from these challenges include the development of deep learning models that significantly outperform traditional segmentation methods. Innovations such as applying convolutional neural networks (CNNs), transfer learning, and ensemble methods have enhanced the sensitivity and specificity of WMH detection (Crimi and Bakas, 2017; Medical Image Computing and Computer Assisted Intervention Society (MICCAI), 2017; Smith and Doe, 2018).

Despite the wide range of proposed solutions, accurate and efficient segmentation of medical images remains a significant problem. While traditional methods are diverse and innovative, they often fail to generalize across different datasets and conditions, primarily due to their reliance on specific intensity and shape features. One significant challenge with current methods, particularly those employing deep learning, is that they are frequently trained on anisotropic data. This type of data, which has different resolutions along different axes, may lead to the development of models that perform well on similar data but struggle with more generalized or heterogeneous datasets. This results in models that exhibit poor generalization capabilities, which makes them less effective in real-world applications where data variability is high. Furthermore, many deep learning models depend on a single data modality, such as MRI, without integrating additional types of medical imaging data that could provide complementary information. This limitation diminishes the reliability and precision of segmentation, as the models cannot fully utilize the comprehensive range of diagnostic information available. While deep learning has markedly advanced medical image segmentation, training on anisotropic data, poor generalization, and reliance on a single data modality continues to present considerable obstacles.

### *1.3 Contributions*

This thesis provides a comprehensive analysis of medical image analysis for White Matter Hyperintensities (WMH), evaluating the performance of Dense UNet, FastSurferCNN, and nn-UNet. It investigates the effects of multimodal information and varying input types and kernel sizes on these models. Additionally, the research explores the impact of transfer learning, using pseudo-labeled data from a large independent cohort to boost model performance. Overall, the thesis not only assesses the capabilities of deep learning architectures in WMH segmentation but also underscores the benefits of integrating multimodal data and applying transfer learning in medical image analysis, focusing on the models' generalizability.

## 2 Material and Methods

### *2.1 Datasets*

In this work, the training and validation datasets are sourced from two population studies: the Rhineland Study (RS)[2] (Breteler et al., 2014; Stöcker, 2016) and the UK Biobank (UKB)[3] (Alfaro-Almagro et al., 2018; Miller et al., 2016), both featuring high-resolution imaging of 1.0 mm.

The first cohort is an ongoing population study from Bonn, Germany, with subjects from 30 years old and above. The study used MR scans recollected by 3 T Siemens MAGNETOM Prisma MRI scanners equipped with 64-channel coils. The core MRI acquisition protocol for every participant in the Rhineland Study includes the following MR contrast: T1w, T2w, Flair, diffusion-weighted, susceptibility-weighted, resting-state functional, and abdominal Dixon MRI with a total net scan time of around 45 minutes. We used the 1 mm Flair, 0.8 mm T1w, and T2w MR scans in this work (Breteler et al., 2014; Stöcker, 2016).

From the Rhineland study using a stratified selection, a random subset ( n = 53 ) consisting of a sex distribution of 70 % of female, 30 % male subjects, minimum age of 32, maximum age of 89, and the average age of

[2] www.rheinland-studie.de
[3] www.ukbiobank.ac.uk

Table 1: Demographics of the stratified selection from Rhineland Study and UK-BioBank participants.

| | Rhineland Study ( n=53 ) | UK-BioBank ( n=623 ) |
|---|---|---|
| **Sex** | | |
| **Women** | 37 (70%) | 330 (53%) |
| **Men** | 16 (30%) | 293 (47%) |
| **Age** | | |
| **Mean (SD)** | 64.8 (13.5) | 64.0 (7.75) |
| **Range** | 32.0 - 89.0 | 45.0 - 83.0 |

64. The UK-Biobank dataset consisted of 49,582 subjects, in Figure 2 shows the white matter hyper-intensity load distribution against the age distribution. In this distribution, we encounter a minimum age of 45, a maximum age of 83, an average age of 64, a female population of 53 %, and a male population of 43 %. To preserve the original balanced distribution of the data was divided into three zones as delimited in Figure 2. Afterward, a sample of a randomized algorithm takes 60 % of the middle area and 20 % of the outer regions. The selection of the data for training ( n = 449 ), validation ( n = 83 ), and test ( n = 91 ). This procedure aims to maintain the same distribution as the original data; the result distribution is shown in Figure 3 maintain the identical distributions as shown in Figure 2.

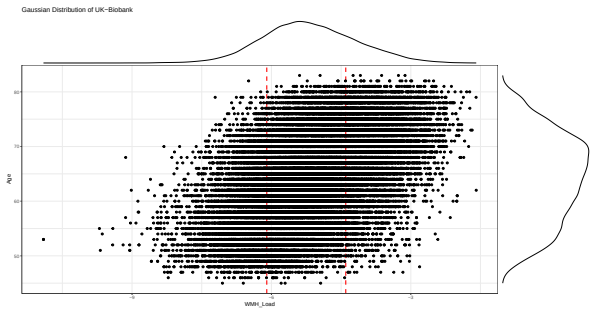



Figure 2: Gaussian Distribution of the UK Biobank dataset with the Age and white matter hyperintensity load against the healthy white matter.

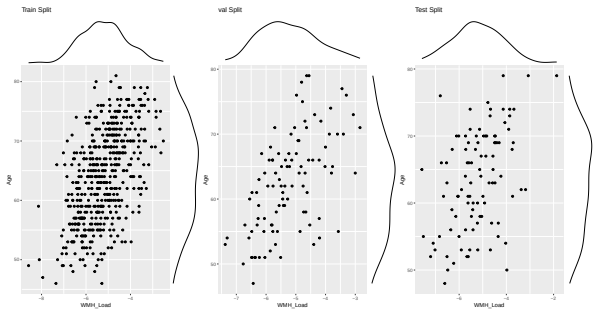



Figure 3: Gaussian distribution of our train, test, and validation data split maintain the same distribution as the whole dataset

To validate the generalizability of the model, samples will be obtained from the *Alzheimer's Disease Neuroimaging Initiative* (ADNI) dataset ( n=5 ) (Alzheimer's Disease Neuroimaging Initiative, 2024) and the whole dataset from MICCAI WMH Challenges of 2017 ( n=52 ) (Kuijf et al., 2022) and 2016 ( n=15 ) (Commowick et al., 2021).

### *2.2 Manual Reference*

This work relies on manual references from five different datasets. In the Rhineland Study, manual annotations were performed on (unprocessed) Flair images by an experienced rater using Freeview, a visualization tool of FreeSurfer (Fischl, 2012; Fischl et al., 2002). For the UK Biobank, utilized the output of the *Brain Intensity AbNormality Classification Algorithm* (BIANCA) (Griffanti et al., 2016) as the manual reference. Despite being an automated method, this output was quality-checked by an expert user (Sundaresan et al., 2022).

The ground truth for the *Alzheimer's Disease Neuroimaging Initiative* (ADNI) (Alzheimer's Disease Neuroimaging Initiative, 2024) dataset and the MICCAI WMH Challenges from 2016 (Commowick et al., 2021) and 2017 (Kuijf et al., 2022) was generated through meticulous manual annotation by expert radiologists. For the ADNI dataset, experts followed standardized procedures to label brain regions and pathologies. In the MICCAI WMH Challenges, multiple experts performed manual segmentations of MRI scans. In cases of discrepancies, consensus methods were employed to ensure reliable ground truth annotations, providing a robust basis for evaluating segmentation algorithms.

### *2.3 Segmentation Networks*

In this work, we used three reliable architectures for segmentation: nn-UNet, Dense UNET, and FastSurferCNN. These architectures were chosen due to their proven effectiveness and reliability in medical image segmentation tasks. Each architecture is composed of four encoder-decoder layers with upsampling and downsampling paths. Their primary difference lies in the feature extractor block, significantly impacting their performance and efficiency. The nn-UNet uses a more traditional UNet structure with optimized hyperparameters, Dense UNET incorporates dense connectivity for improved feature propagation, and FastSurferCNN leverages a streamlined architecture optimized for speed and accuracy. These distinctions and their implications will be discussed in detail in this section.

The Dense UNet incorporates the use of a *Dense Block*, denoted as DB in Figure 4. This block consists of four layers connected sequentially, where the input to each layer includes the feature maps from all preceding layers as shown in equation 1. This connectivity is achieved through convolutions, commonly known as *dense connections*. Each layer in the dense block is represented by the operation $H$, which includes batch normalization, *Parametric ReLU*, and a 3x3 convolution. After extensive experimentation with different kernel sizes, including 5x5, the 3x3 convolution provided

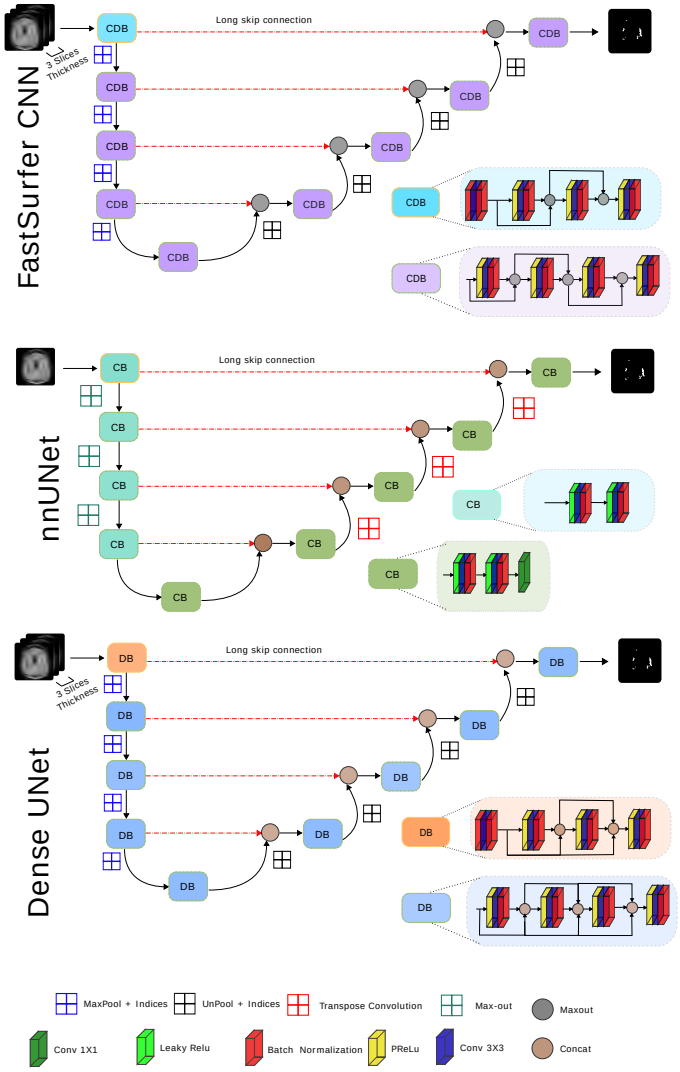



Figure 4: From top to bottom: Fast-Surfer CNN, nn-UNet and Dense UNet architectures. Input modalities are processed before being ingested into the models. Note: Competitive Dense Block (CDB), Dense Block (DB) and Convolutional Blocks (CB).

better performance, as documented in Table 3. These dense connections facilitate better gradient flow during training, which helps mitigate the vanishing gradient problem. Additionally, they promote feature reuse by allowing layers to access features from earlier layers, and they enhance parameter efficiency by reducing the need for redundant filters. These advantages contribute to the Dense UNet's improved performance in segmentation task (Jégou et al., 2017; Safarov and Whangbo, 2021).

$$x_3 = H_3([x_3, x_2, x_1, x_0]) \quad (1)$$

The FastSurferCNN (Henschel et al., 2020) introduces a variation of the Dense Block with a max-out operation to select the most relevant features. This new block, called the *Competitive Dense Block* (CDB), is depicted in Figure 6. The maxout function reduces computational complexity and prevents the network from being overwhelmed by redundant information, thereby propagating only the most significant features through the network. This architecture enhances efficiency and performance by optimizing the model's capacity to learn meaningful representations (Estrada et al., 2018).

The nnU-Net (Isensee et al., 2021) architecture employs *Convolutional Blocks* (CB) with batch normalization and activation functions, using max-pooling for downsampling and transposed convolutions for upsampling. This combination of elements ensures that the network can handle various segmentation tasks with high accuracy and efficiency. nnU-Net's automatic configuration process, which covers the entire segmentation pipeline from preprocessing to post-processing, allows it to adapt to new datasets with minimal manual intervention, making it an ideal foundation for developing advanced segmentation models.

The nnU-Net framework is particularly valuable due to its holistic approach to segmentation pipeline con-

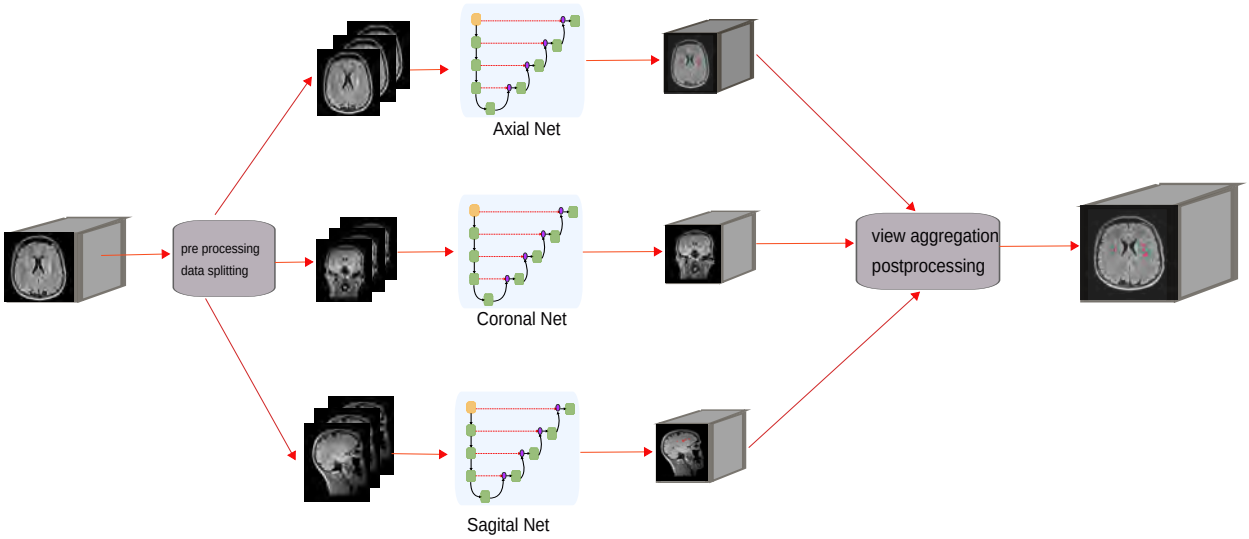



Figure 5: Proposed pipeline for WMH segmentation. The pipeline is divided into three stages: First, preprocessing and data reorganization. Then, WMH tissue segmentation within each volume plane, and finally, an ensemble of predicted label maps.

figuration. It systematically addresses the challenges of designing and optimizing deep learning-based segmentation methods by using a set of fixed, rule-based, and empirical parameters. This automated configuration process enables nnU-Net to generalize well across diverse datasets, outperforming many specialized methods. As an out-of-the-box tool, nnU-Net simplifies the deployment of state-of-the-art segmentation techniques, making them accessible to a broader audience without the need for extensive expertise in deep learning or computational resources beyond standard network training (Isensee et al., 2021).

### *2.4 Model learning*

The harmonization of data is ensured through the implementation of specific pre-processing steps. This is necessary given that the data in question has different specifications. Initially, the images are interpolated to a 1 mm spatial resolution, after which they are aligned to the RAS[4] orientation. The objective of this standardization process is to guarantee consistency in the physical space and orientation of the MRI scans. Subsequently, the images are formatted in accordance with the specifications of the deep learning algorithm, which requires an input size of *256x256x256*. Rather than applying the conventional *MinMax* rescaling method, we employ a percentile method for scaling MRI intensities, which effectively redistributes the intensities between 0 and 255.

Furthermore, our preprocessing procedure entails appending a data package of three slices, both before and after the target image. This enhances the spatial context and intensity information, which are critical for accurate segmentation. To mitigate the creation of numerous zero maps from our lesion segmentation map, a strategy was implemented whereby only 10 % of the total non-lesion samples of the MRI volume were selected. Our labels exhibited a significant imbalance in the number of instances per class due to the limited number of lesions relative to the volume.

As a result, a corrective mechanism is needed. Roy et al. (2017) suggested calculating weights for each label map with the aim of improving the propagation of losses. This approach specifically aims to increase the emphasis on detecting small lesions by assigning greater weight to them in the loss propagation process.

$$\omega(x) = \sum_{l} I(S(x) = l) \frac{\text{median}(f)}{f_l} + \omega_0 \cdot I(\|\nabla S(x)\| > 0) \tag{2}$$

Equation 2 tailors the loss function to challenges arising from the unbalanced label map and the error in the anatomical boundaries. The first term models the median frequency balancing and compensates for the class imbalance problem by highlighting classes with low probability. The second term puts higher weight on anatomical boundary regions to emphasize the correct contour segmentation; at last, the term $\omega_0$ balances the two terms (Roy et al., 2017).

In magnetic resonance imaging, the three-dimensional representation of lesions significantly influences segmentation outcomes due to varying characteristics across different reference planes. The voxel size exerts a significant influence on the data volume for each plane. Moreover, the computational demands for 3D segmentation are considerably more substantial than those for 2D segmentation. Our approach, inspired by the ensemble method of three networks trained on three anatomical representations (Park et al., 2021), is illustrated in our pipeline in Figure

[4]Right - Anterior - Superior

5. The pipeline employs three networks to generate a view-aggregation, whereby the probability maps generated by each network's predictions are reoriented to the sagittal plane. A new model is then generated by selecting the maximum value from the 2D projections predictions. A final segmentation is then produced by applying a threshold of 0.9. This method effectively converts two-dimensional data into a comprehensive three-dimensional representation, capitalizing on the strengths of multi-plane analysis.

Throughout the training phase, these configurations are selected: an Adam optimizer with a Base Learning Rate of 0.01, a weight decay of $1 \times 10^{-4}$, betas between 0.9 - 0.999, and an eps of $1 \times 10^{-8}$. On the other hand, we decided to use a multistep configuration with a milestone of 70 and a gamma of 0.1. The scheduler helps us to apply a decay on the learning rate at milestone 70, maintaining a stable learning rate after this happens. Since we are limited in resources, a super epoch methodology was implemented, assuring that every 16 epochs, the backpropagation occurred; this ensures the same behavior no matter the hardware constraints.

The work from Yeung et al. (2022) explores the impact of different losses on class imbalance medical image segmentation. In which, a combination of Distribution-based and Region-based losses reach above-average results on the testing data set; therefore, we intend to recreate a similar loss in this work. Cross-entropy loss was chosen in the distribution part due to the ability to measure the difference between two probability distributions for a given random variable, minimize pixel-wise error, and use weights to influence the prediction area. Given a binary classification of our problem of lesion and background $c \in \{1, 2\}$, the loss function for a single sample is defined as:

$$\mathcal{L}_{\text{CCE}} = -\sum_{c=1}^{2} w_c y_{n,c} \log\left(\frac{\exp(x_{n,c})}{\sum_{i=1}^{2} \exp(x_{n,i})}\right) \tag{3}$$

In which $x_{n,i}$ represents the raw output (logit) of the model for the nth sample and class i, $y_{n,c}$ is one of the nth sample is of class c and 0 otherwise and $w_c$ is the weight associated with class c, used to give more or less importance to the class in the loss computation.

The Region-based loss to use is the Dice Loss, which measures the intersection of pixels with a sample we are evaluating. This loss can be adapted to handle class imbalance. However, this loss is inherently unstable where there is highly class-imbalanced data. The formula used in this project is shown below:

$$\mathcal{L}_{\text{DSC}} = 1 - \frac{2TP}{2TP + FP + FN} \tag{4}$$

The combo loss is simple a combination of the two losses, but we added some weights to play around the influence of each of the losses during our training.

$$\mathcal{L}_{\text{total}} = \omega_{\text{DSC}} \cdot \mathcal{L}_{\text{DSC}} + \omega_{\text{CCE}} \cdot \mathcal{L}_{\text{CCE}} \tag{5}$$

In order to create more diversity from our data ingested into the model, a data augmentation pipeline was implemented into the data loader. Table 2 summarizes the complete augmentation list with the principal parameter used in each augmentation. In order to achieve this pipeline, we used *torchio* [5] library since we can apply the data augmentation on the fly, which decreases the use of memory. The use of this augmentation serves to mimic real scenarios while acquiring images from real-world data. The data usually does not have the same visual point. In order to surpass this, we change the zooming of the images, the direction of the images, and move the center of the image. The Gaussian noise and Bias Field add external distortions such as those produced by the MRI or pixel-wise noises from the sensors.

Table 2: Summary of Image Transformations. Note: The percentage of appearance was divided between all of them in order to choose one per 50%

| Transformation | Parameters | Description |
|---|---|---|
| **Geometric Transformations** | | |
| Scales | `scales=(0.8, 1.15)` | Isotropic scaling |
| Rotation | `degrees=10` | Rotation within 10 degrees |
| Angle Flip | `degrees=180` | Flipping the image |
| Translation | `translation=(15.0, 15.0, 0)` | Translating the image |
| **Intensity Transformations** | | |
| Bias Field | `coefficients=0.5` | Random bias field adjustment |
| Gaussian Noise | `std=(0.01, 0.1)` | Adding noise with specified std deviation |

### 2.5 *Metrics*

To evaluate the segmentation outcomes, we selected four different metrics to assess the similarity between our predicted label map and manual annotations. The first aspect of evaluating the spatial overlap consensus with the dice similarity coefficient (DICE) (Taha and Hanbury, 2015). Let M (manual annotations) and P (prediction) denote binary label maps, then Dice is defined as:

$$Dice = \frac{2 \cdot |M \cap P|}{|M| + |P|} \tag{6}$$

The numerator part represents the elements that match, and the denominator represents the sum of all the elements in the manual and prediction map. The second aspect is how much volumetric similarity (Taha and Hanbury, 2015) it is with the annotation map. VS is defined as:

$$VS = 1 - \frac{||M| - |P||}{||M| + |P||} \tag{7}$$

The third metric evaluated the quality of the segmentation boundary delineation (contour). In this work, we choose a 95% Hausdorff distance (HD) (Taha and Hanbury, 2015) as it is less sensitive to outliers. HD95 is

[5] https://torchio.readthedocs.io/

considered as the 95th percentile of the ordered distance measures, and it is defined as:

$$d_{95}(M, P) = 95^{\text{th}}\{m \in M \min_{p \in P} d(m, p)\} \quad (8)$$

$$d_{H95}(M, P) = \max(d_{95}(M, P), d_{95}(P, M)) \quad (9)$$

Finally, we will use some measurements to detect the correct lesion since our model could generate a false negative or over-segmentation. In order to measure this, we are using Recall (Taha and Hanbury, 2015) as the sensitivity and the average of precision, defined as:

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

To confirm statistically significant differences in segmentation performance, we employed the Wilcoxon signed-rank, non-parametric paired test (Wilcoxon, 1992). This rigorous statistical method guarantees the robustness and reliability of our results. In addition, we used a comprehensive ranking approach to compare the performance of different segmentation models. We first ranked the performance of each model for individual metrics and then calculated an overall rank by taking the geometric mean of these individual rankings.

## 3 Experiments and Results

### *3.1 Ablation Analysis*

In this section, an ablation analysis was conducted to identify the optimal network configuration and input modalities configuration.

#### *3.1.1 Kernel Analysis*

In the experimental section, we perform a comprehensive kernel analysis to determine the optimal kernel size for our segmentation models. The findings, derived from ablation tests conducted without data augmentation and using consistent loss functions as described in Section 2.4 and also the same loss function in each of the ablation tests. Table 3 presents the results, which indicate a significant influence of kernel size on model performance. This data drove our decision to adopt a 3x3 kernel across all networks, aiming to achieve the best balance between accuracy and computational efficiency.

#### *3.1.2 Input Analysis*

This section delves into the differential spatial information provided by various MRI modalities as depicted in Figure 6. Given the distinct data characteristics each modality offers, it becomes imperative to assess the individual and combined effects of these modalities on model performance. For this analysis, the inputs considered include: Flair, T1, T2, Flair-T1, Flair-T2, and the combination of Flair-T1-T2. All findings reported stem from the validation set, with the training benefitting from the data augmentation routines detailed in Section 2.

Table 3: Mean (and standard deviation) segmentation performance of the validated models on validation-set with the 3x3 or 5x5 kernel for the models convolutions. The best metric per model is show in bold.

| Model<br>Flair input | Dice ↑<br>Mean (SD) | VS ↑<br>Mean (SD) | Recall ↑<br>Mean (SD) |
|---|---|---|---|
| Dense UNET | | **3x3** | |
| | **0.6679 (0.1835)** | **0.8537 (0.1039)** | **0.7716 (0.1521)** |
| | | **5x5** | |
| | 0.6227 (0.2126) | 0.8155 (0.1301) | 0.7497 (0.1776) |
| FastSurferCNN | | **3x3** | |
| | **0.6311 (0.2017)** | **0.9157 (0.0702)** | **0.7957 (0.1502)** |
| | | **5x5** | |
| | 0.6070 (0.2102) | 0.7493 (0.1596) | 0.7981 (0.1627) |

In Table 4, all the metrics for the input analysis of the ablation analysis are presented. The literature explored during the section 1.2 presented a current trend in using the Flair images as the gold standard for the segmentation of this lesion, as shown in the table 4 from the single input test is the one that yields the highest values in all the metrics. Conversely, the trend also shows that using two inputs benefits the model using information from 2 different space domains.

In the multiple inputs, it is shown that it helps in some of the networks to get a higher dice, but on the other hand, some of the combinations yield a higher Hausdorff distance. The other metrics maintain similar to in trend with the single input distance. At last, using the 3 spatial domains available in our dataset gets us a a higher number in all the metrics, however this come with a high computational cost since model is slower and requires more memory. Since, the UK-BioBank dataset had available only 2 modalities, we choose the Flair and Flair-T1 to implement in our testing.

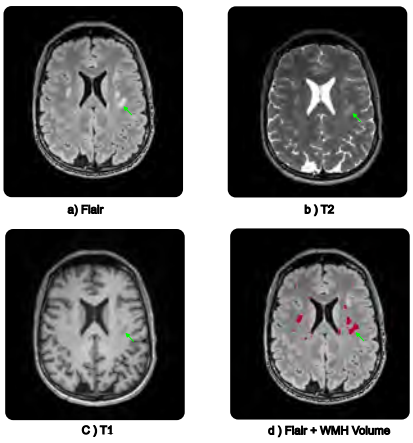



Figure 6: MRI modalities available on the Rhineland Study dataset. From top to bottom: Flair, T2, T1 and Flair+Lesions Volume (in red). Note: The green arrow points the same lesion in all the features spaces, with a high information difference between them.

Figure 8 shows the qualitative results from the segmentation algorithm using only the Flair modality, the results from left to right: Original MRI image (Flair), manual annotation map (GT), FastSurferCNN, Dense UNet, and nnU-Net. The rows represented by *a*, *b*, *c*, and *d* show four different samples from different cases of the test set. The arrows in samples *a* and *b* represent an over-segmentation from the algorithm where two close different regions are fused, causing a higher volume with a higher number of false positives. Conversely, the *c* and *d* are examples of missing segmentation since the models can not label these pixels correctly or only recognize some parts of the manually annotated map, causing a higher number of false negatives.

Furthermore, Figure 9 shows the qualitative results from dual-modality, the Flair and T1, training in the unseen dataset with the same distribution as Figure 8, with different case from the rows. The data was ingested in the model using double channels and concatenation of both feature maps. The *b* represents some cases where the model fails to recognize the lesion, even in the dual modality-specific feature maps. In contrast, in cases *a* and *d*, all the models fail to recognize the some small lesion, which has impacted all the metrics. Finally, in case *c*, we observed a general failure in all the models to segment the tiny lesion, and their predictions over this layer were consistent. In this last case, this brain can be cataloged as an out-of-the-distribution patient; more of this is discussed in Section 4.

Table 4: Mean (and standard deviation) segmentation performance of the validated models on validation-set of Rhineland Study with all the input combination. The overall best metric is shown in bold.

| Model | Dice ↑ Mean (SD) | VS ↑ Mean (SD) | HD95 ↓ (mm) Mean (SD) | Recall ↑ Mean (SD) |
|---|---|---|---|---|
| **Flair** | | | | |
| Dense UNET | 0.7046 (0.1953) | 0.9038 (0.1424) | 8.6554 (7.3333) | 0.6833 (0.2320) |
| FastSurferCNN | 0.7206 (0.1581) | 0.9238 (0.0605) | 7.8664 (6.3261) | 0.7345 (0.1925) |
| nn-Unet | **0.7223 (0.1925)** | 0.8131 (0.2327) | **6.7714 (5.1322)** | 0.6442 (0.2316) |
| **T1** | | | | |
| Dense UNET | 0.5373 (0.2066) | 0.8756 (0.0978) | 12.6140 (6.2992) | 0.5947 (0.1872) |
| FastSurferCNN | 0.5446 (0.1976) | 0.8065 (0.1329) | 12.8194 (6.3562) | 0.6648 (0.1681) |
| nn-Unet | 0.5142 (0.2790) | 0.7217 (0.2859) | 15.1769 (14.2329) | 0.4479 (0.2849) |
| **T2** | | | | |
| Dense UNET | 0.5887 (0.2033) | 0.9302 (0.0389) | 16.7864 (12.0687) | 0.5913 (0.1906) |
| FastSurferCNN | 0.5873 (0.2092) | 0.8778 (0.0827) | 16.4945 (12.6525) | 0.6365 (0.1809) |
| nn-Unet | 0.5641 (0.2827) | 0.7257 (0.3308) | 14.8174 (16.2419) | 0.4964 (0.2719) |
| **Flair-T1** | | | | |
| Dense UNET | 0.6981 (0.1489) | 0.8245 (0.1747) | 14.3157 (10.618) | 0.7337 (0.1917) |
| FastSurferCNN | 0.6902 (0.1515) | 0.8290 (0.1405) | 14.1490 (11.157) | 0.7623 (0.1900) |
| nn-Unet | 0.7083 (0.2079) | 0.8093 (0.2525) | 7.4703 (9.9173) | 0.6351 (0.2430) |
| **Flair-T2** | | | | |
| Dense UNET | 0.7078 (0.1582) | 0.9334 (0.0283) | 8.9039 (5.7597) | 0.7388 (0.1715) |
| FastSurferCNN | 0.6997 (0.1627) | 0.8984 (0.0603) | 9.8864 (5.9643) | **0.7758 (0.1640)** |
| nn-Unet | 0.7073 (0.1981) | 0.7961 (0.2240) | 8.2565 (8.0085) | 0.6202 (0.2256) |
| **Flair-T1-T2** | | | | |
| Dense UNET | 0.7225 (0.1503) | **0.9386 (0.0223)** | 7.7507 (8.0966) | 0.7542 (0.1376) |
| FastSurferCNN | 0.7206 (0.1503) | 0.9183 (0.0378) | 7.8753 (8.0953) | 0.7748 (0.1427) |
| nn-Unet | 0.7161 (0.1874) | 0.8030 (0.2215) | 8.3508 (8.8113) | 0.6300 (0.2202) |

### *3.2 Performance Analysis*

The section present the result of the best models, after the decision took on the ablation test carry on the Section 3.1, using the full training routine and the unseen test dataset from Rhineland Study. Finally, we present a robust analysis using the optimal model identified in the previous performance analysis to assess its generability to previously unseen datasets and different data domains.

#### *3.2.1 Transfer Learning*

Even dough segmentation performance using only Rhineland Study data shows promising results. The Rhineland labeled subset lacks a demographic diversity. Therefore to fill the gap in the RS distribution, we proposed utilizing the Uk Biobank labels for initializing the models. The UK Biobank labels are obtained from a automated method therefore, missing distribution can be completed by choosing a stratified subset.

The result of training the models with the UK Biobank are shown in Table 6; the segmentation performance in the unseen data shows that the models pick up good relationships from the data, classification with huge training dataset is achievable, and the segmentation from this training is optimal. The results only reflect the training needed to apply the transfer learning; at this point, these results are not comparable with the ones trained with the RS.

On the other hand, we evaluated the impacts of using the fine-tuning methodology in comparison to using the knowledge graphs from only learning from one dataset, Figure 7 shows some examples of these impacts. From top to bottom, we show the same two participants from the unseen data set from the one and two modalities training. The figure used only two algorithms that were evaluated with the transfer learning. Figure 7 shows the overlay of the ground-truth map with each of the prediction's maps. Therefore, we can observe the actual impacts of using this methodology with a dataset with a normal distribution. The effects of this methodology are marked with the use of a green and blue arrow, the improvements in helping distinguish between false positives, and the inclusion of their feature map false positives are shown respectively.

#### *3.2.2 Standalone Segmentation*

This section, explores the segmentation capabilities of the best models using Flair and the combination of Flair with T1 MRI modalities. Table 5, we present the training results using two modalities available; in the table, we also see the result of the overall ranking with the quadratic mean, obtained after individually ranking all the metrics used to compare the models. Furthermore, Table 5 shows that the paired Wilcoxon test over VS, HD95 and DICE does not yield any value with a statistically significant impact; on the other hand, in the metric Recall, we observed a lot of statistical significance between the test cases.

The comparison of all the models, illustrated thought box plots in the Appendix A.10 for the metrics dis-

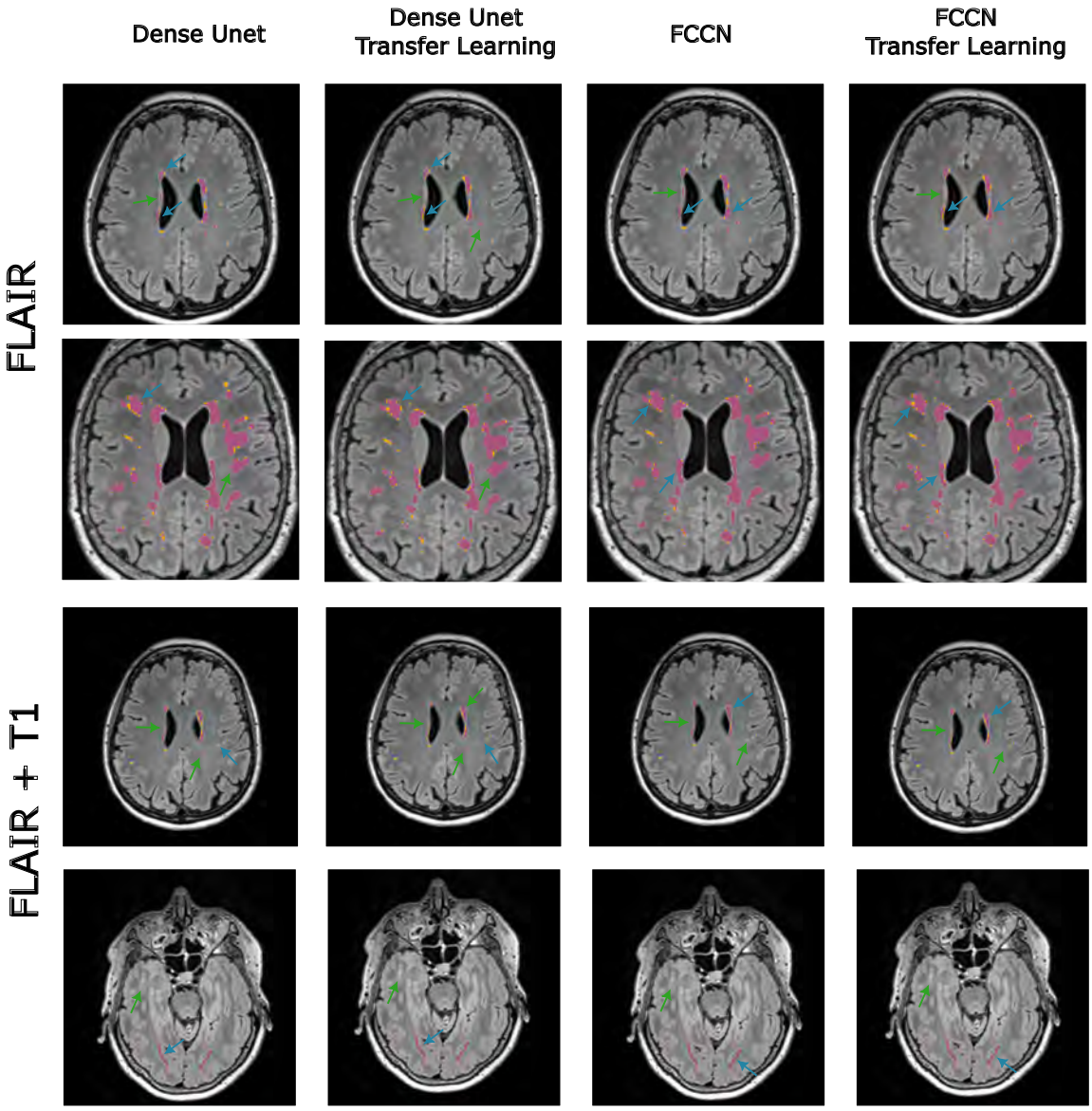



Figure 7: Comparison of the normal vs. transfer learning predictions from the Dense UNet and FCCN for two participants of the in-house test-set on Flair and T1 input. (A-D). The figure show in orange the GT with a overlay in brown of the prediction. The structures pointed by the green arrow represent improvements and the blue one represents deterioration. Note: each row represents a different participant with corresponding MRI modalities (Flair and T1) and automated generated segmentation on the axial plane.

Table 5: Mean (and standard deviation) segmentation performance of the validated models on the unseen test-set

| Model | Overall Rank | Dice ↑ Mean (SD) | VS ↑ Mean (SD) | HD95 ↓ (mm) Mean (SD) | Recall ↑ | Signif. Mean (SD) |
|---|---|---|---|---|---|---|
| **Flair** | | | | | | |
| a: Dense UNET | 2.63 | 0.6728 (0.1190) | 0.8521 (0.1122) | 8.5841 (8.6607) | 0.7203 (0.1506) | b |
| b: FastSurferCNN | 2.24 | 0.6681 (0.1297) | **0.8583 (0.1587)** | 8.9319 (9.3002) | **0.7652 (0.1353)** | d,e |
| c: nn-Unet | 2.24 | **0.6898 (0.1303)** | 0.8287 (0.1451) | **6.9278 (8.7652)** | 0.7006 (0.1769) | - |
| d: FastSurferCNN (TF) | **3.46** | 0.6804 (0.1161) | 0.8483 (0.1155) | 8.3087 (9.3607) | 0.7338 (0.1501) | e |
| e: Dense UNET (TF) | 2.63 | 0.6778 (0.1158) | 0.8464 (0.1090) | 8.2911 (9.5354) | 0.7060 (0.1540) | - |
| **Flair+T1** | | | | | | |
| a: Dense UNET | **3.66** | 0.6667 (0.1332) | **0.8647 (0.1567)** | 14.1827 (17.8801) | 0.7354 (0.1404) | b |
| b: FastSurferCNN | 2.51 | 0.6571 (0.1428) | 0.8397 (0.1598) | 19.1786 (30.3402) | **0.7674 (0.1390)** | c,d,e |
| c: nn-Unet | 2.51 | 0.6692 (0.1281) | 0.8315 (0.1497) | **7.6416 (9.5324)** | 0.6514 (0.1722) | - |
| d: FastSurferCNN (TF) | 2.06 | 0.6606 (0.1376) | 0.8383 (0.1582) | 22.5771 (33.2519) | 0.7335 (0.6606) | e |
| e: Dense UNET (TF) | 2.51 | **0.6722 (0.1239)** | 0.8287 (0.1451) | 9.7992 (10.5880) | 0.7111 (0.1663) | - |

Note: the statistical significance column (Signif.) indicates which other models the model outperforms (paired Wilcoxon signed-rank test, $p < 0.05$). Overall Rank is the Quadratric mean after ranking all the evaluation metrics.

Table 6: Mean (and standard deviation) segmentation performance of the validated models on an unseen dataset of UK-Biobank Study with all the input combinations. Best metric value per model is shown in bold. Note: These results cannot be compared with the Rhineland Study.

| Model | Dice ↑ Mean (SD) | VS ↑ Mean (SD) | HD95 ↓ (mm) Mean (SD) | Recall ↑ Mean (SD) |
|---|---|---|---|---|
| **Flair** | | | | |
| Dense UNET | 0.7567 (0.1108) | 0.8847 (0.1040) | 3.7385 (3.2790) | 0.8030 (0.1400) |
| FastSurferCNN | 0.7556 (0.1073) | 0.8738 (0.1028) | **3.5336 (3.0265)** | **0.8333 (0.1290)** |
| nn-Unet | **0.7574 (0.1009)** | **0.9061 (0.0849)** | 4.5323 (3.4513) | 0.7733 (0.1389) |
| **Flair-T1** | | | | |
| Dense UNET | 0.7640 (0.1173) | 0.8656 (0.1204) | 3.2099 (2.9686) | 0.8158 (0.1475) |
| FastSurferCNN | 0.7591 (0.1160) | 0.8575 (0.1218) | **3.0361 (2.8058)** | **0.8448 (0.1291)** |
| nn-Unet | **0.7734 (0.1061)** | **0.8973 (0.0999)** | 4.7176 (3.9876) | 0.7577 (0.1527) |

cussed in Section 2: Dice coefficient, Volume Similarity (VS), Hausdorff Distance (HD95), and Recall. The algorithms compare for WMH segmentation with Dense U-Net, FastSurferCNN, and nnU-Net, each applied to different imaging modalities like Flair and Flair-T1, with a color-coded scheme to distinguish between them. The Appendix A.10, the values for the Dice coefficient range from approximately 0.4 to 0.9, indicating varying levels of segmentation accuracy across algorithms, assuring a high pixel match with the GT. Volume Similarity values are predominantly near 0.9, suggesting a high volumetric consistency. Hausdorff Distance varies, with measurements extending from nearly 0 mm to about 15 mm, reflecting differences in boundary precision. Recall scores span from 0.4 to 0.9, showing the varied effectiveness of algorithms in identifying relevant instances. In the displayed box plots, outliers are evident across several metrics, signaling occasional deviations in algorithm performance. Specifically, Volume Similarity presents many outliers, indicating significant deviations from typical volume agreements. Lastly, Recall demonstrates outliers mainly in the lower range across various algorithms, indicating failures in correctly identifying a higher proportion of actual positives. These outliers highlight potential weaknesses or limitations of the algorithms under challenging conditions or with unusual data types.

As shown in Table 5, we observed that the best overall rank model in only one modality input is FastSurferCNN (TF) since it obtained a balanced performance across all metrics. nn-Unet demonstrated superior Dice scores (0.6892 ± 0.1268), suggesting better segmentation overlap. FastSurferCNN also exhibited the highest VS (0.8583 ± 0.1587) and nn-UNet significantly outperformed others in HD95 (6.5882 ± 8.4740), indicating precise boundary segmentation. FastSurferCNN achieved the highest recall (0.7652 ± 0.1325), suggesting fewer false negatives. For the Flair-T1 input, Dense UNET ranked highest overall (3.66), followed by nn-UNet (2.66). Regarding Dice scores, nn-UNet (0.6755 ± 0.1187) performed best. Dense UNET exhibited the highest VS (0.8647 ± 0.1567), while nn-UNet again demonstrated the best boundary precision with the lowest HD95 (7.1548 ± 7.6686). FastSurferCNN achieved the highest recall (0.7674 ± 0.1390), reinforcing its robust detection capability. Overall, the nn-UNet model stands out for its exceptional boundary precision (HD95) in both Flair and Flair-T1 inputs. FastSurferCNN and FastSurferCNN (TF) consistently perform well across multiple metrics, particularly in recall, indicating strong detection capability. nn-Unet performed well in Dice and VS but is less consistent in recall. Dense UNET (TF) benefits from the additional T1 modality, particularly in VS. The incorporation of T1 data generally enhances performance metrics, underscoring the value of multimodal inputs for improved segmentation performance.

### *3.2.3 Generability Analysis*

This subsection, we evaluate the robustness and applicability of different best models from the Section 3.2 across various datasets within the context of medical imaging for WMH Segmentation. Table 7 presents the mean (and standard deviation) segmentation performance of validated models on an unseen dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI), MICCAI 2016 and MICCAI 2017 using Flair input. The models that were evaluated include Dense

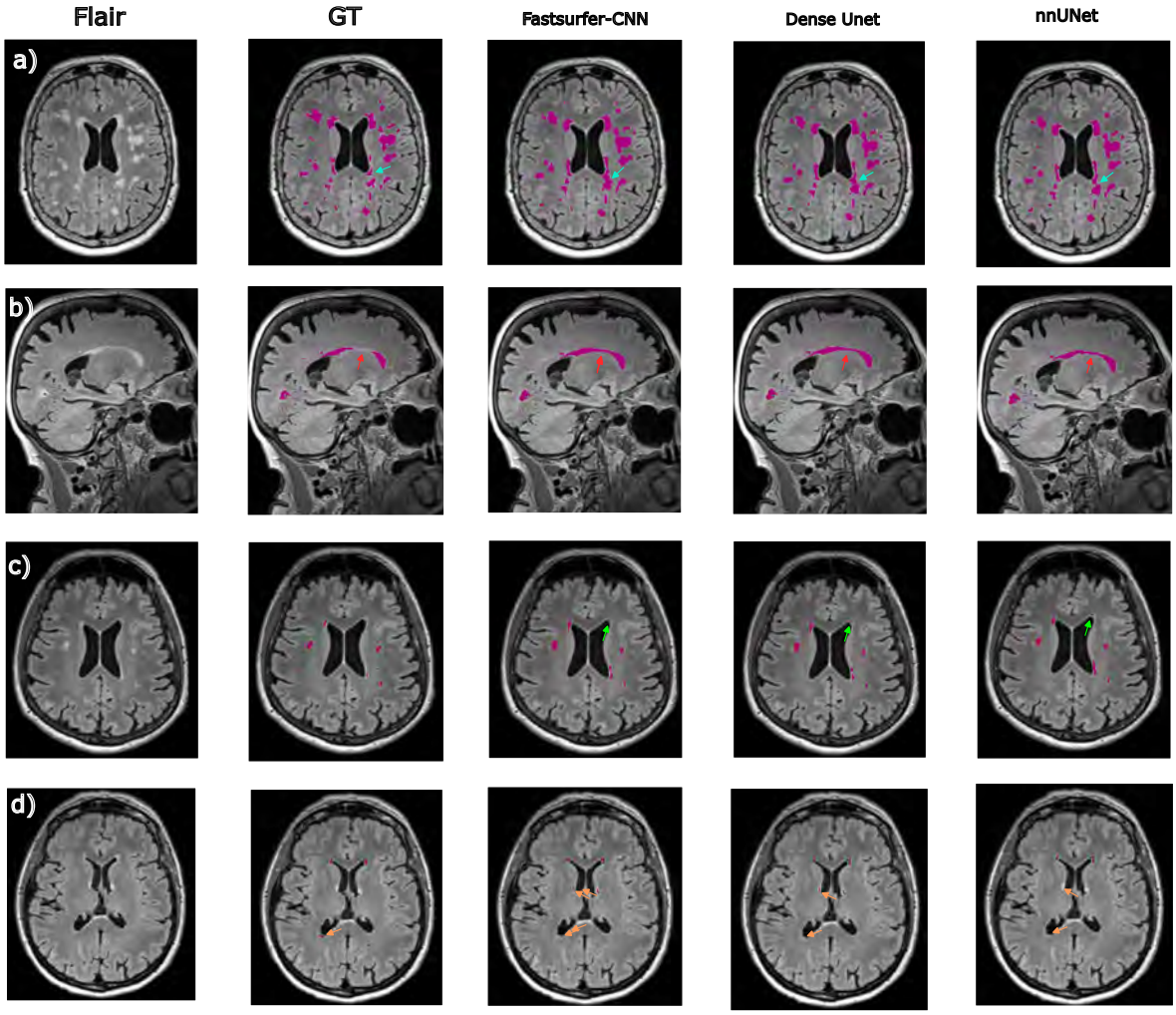



Figure 8: Comparison of the ground Truth vs. predictions from the proposal method and benchmarks for four participants of the in-house test-set on Flair input. (A-D). All automated methods generate similar segmentation to the manual ones. However, some differences are observed in some on some of the structures pointed by the arrows. Note: each row represents a different participant with corresponding MRI modality (Flair), manual ground truth (GT), and automated generated segmentation on the axial and saggital plane.

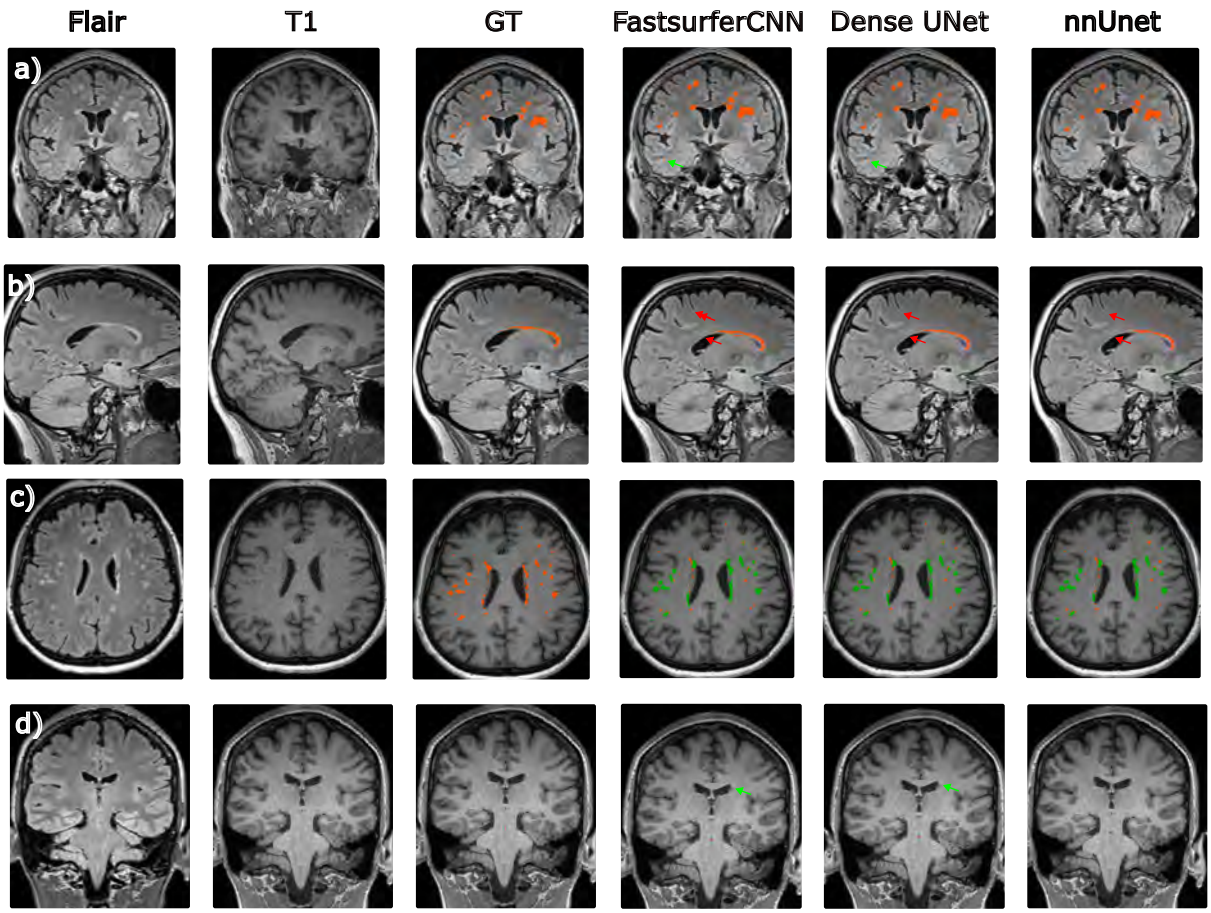



Figure 9: Comparison of the ground Truth vs. predictions from the proposal method and benchmarks for four participants of the in-house test-set on Flair and T1 input (A-D). All automated methods generate similar segmentation to the manual ones. However, some differences are observed in some on some of the structures pointed by the green arrow represent false positives and the red one represents false negatives. Note: each row represents a different participant with corresponding MRI modalities (Flair and T1), manual ground truth (GT), and automated generated segmentation on the axial, coronal and sagittal plane.

Table 7: Mean (and standard deviation) segmentation performance of the validated models on an unseen dataset of ADNI, MICCAI 2016 and MICCAI 2017 All measurements are in percentual value and only Hausdorff Distance is in millimeters.

| Model | Dice ↑ Mean (SD) | Signif. | VS ↑ Mean (SD) | Signif. | HD95 ↓ Mean (SD) | Signif. | Recall ↑ Mean (SD) | Signif. |
|---|---|---|---|---|---|---|---|---|
| **ADNI Dataset (n = 5)** | | | | | | | | |
| a: Dense UNET | 0.4978 (0.2307) | - | 0.6214 (0.1166) | - | 14.8668 (11.5912) | - | 0.3732 (0.2002) | - |
| b: Dense UNET (TF) | 0.5116 (0.2512) | - | 0.6740 (0.1043) | - | 17.3694 (16.0344) | - | 0.4253 (0.1699) | - |
| c: FastSurferCNN | 0.5121 (0.2424) | - | 0.7155 (0.1143) | - | **13.3401 (8.9113)** | - | **0.8333 (0.1290)** | - |
| d: FastSurferCNN (TF) | **0.5748 (0.2706)** | - | **0.7882 (0.1865)** | - | 16.3177 (15.9769) | - | 0.5430 (0.1786) | - |
| e: nn-Unet | 0.0375 (0.0444) | - | 0.0472 (0.0565) | - | 33.9479 (1.2770) | - | 0.0197 (0.0375) | - |
| **MICCAI 2016 Dataset ( n = 15 )** | | | | | | | | |
| a: Dense UNET | 0.6048 (0.1848) | - | 0.7711 (0.1824) | - | 14.1606 (8.4260) | e | 0.6381 (0.2002) | d,e |
| b: Dense UNET (TF) | 0.6048 (0.1680) | - | 0.7749 (0.1580) | - | 14.2158 (8.3081) | e | 0.6463 (0.1743) | e |
| c: FastSurferCNN | **0.6102 (0.1861)** | - | 0.7768 (0.1823) | - | 13.5418 (8.6273) | e | 0.6536 (0.1747) | a,d,e |
| d: FastSurferCNN (TF) | 0.6095 (0.1820) | - | **0.7872 (0.1766)** | - | **13.2680 (8.9691)** | b,e | **0.6870 (0.1632)** | b,e |
| e: nn-Unet | 0.5608 (0.1189) | - | 0.6940 (0.1697) | - | 17.2830 (7.0809) | - | 0.4427 (0.1347) | - |
| **MICCAI 2017 Dataset ( n = 52 )** | | | | | | | | |
| a: Dense UNET | 0.5853 (0.1899) | b,d,e | 0.7467 (0.1824) | b,d,e | 15.5831 (11.3592) | b,d,e | 0.7346 (0.1712) | b,d,e |
| b: Dense UNET (TF) | 0.5220 (0.1975) | e | 0.6781 (0.2207) | e | 18.9549 (12.8932) | e | 0.7404 (0.1631) | - |
| c: FastSurferCNN | 0.5833 (0.1891) | e | 0.7306 (0.1814) | a,b,d,e | 15.8757 (11.4957) | b,d,e | 0.7523 (0.1676) | a,b,d,e |
| d: FastSurferCNN (TF) | 0.5225 (0.2012) | e | 0.6530 (0.2195) | b,e | 19.0830 (12.8639) | e | 0.7758 (0.1683) | b,e |
| e: nn-Unet | **0.6023 (0.1725)** | - | **0.8087 (0.2071)** | - | **11.4667 (9.0187)** | - | 0.5601 (0.2099) | - |

Note: the statistical significance column (Signif.) indicates which other models the model outperforms (paired Wilcoxon signed-rank test, $p < 0.05$). Best performances are highlighted in bold.

UNET, Dense UNET (TF), FastSurferCNN, FastSurferCNN (TF) and nn-Unet with Flair input. The following performance metrics were used: Dice coefficient, Volume Similarity, Hausdorff Distance at 95 % and Recall.

The ADNI dataset demonstrates that models exhibit moderate Dice scores. Of these models, the FastSurferCNN (TF) model achieves the highest Dice score (0.5748). The FastSurferCNN (TF) model also exhibits strong performance in the VS and Recall metrics, although it is not the model with the highest HD95 score, which the Dense UNET model achieves. In the MICCAI 2016 dataset, the performance of models improved in general. The FastSurferCNN model demonstrated exceptionally high Dice and the FastSurferCNN (TF) VS metric scores. In the larger MICCAI 2017 dataset, the Dense UNET model performs exceptionally well in the Dice metric, surpassing other models significantly. This is indicated by its statistical outperformance of models b, d, and e.

The nn-UNet model demonstrates highly variable performance across different medical imaging datasets, as indicated by its evaluation of the ADNI, MICCAI 2016, and MICCAI 2017 datasets. Initially, the model exhibits inferior performance on the ADNI dataset with minimal segmentation overlap, as evidenced by poor Dice and recall scores. However, there is a notable enhancement in performance on the subsequent MICCAI datasets, with Dice scores indicating more significant overlap and HD95 scores indicating reduced boundary discrepancies. Despite these improvements, it shows that the overall variability in metrics such as Dice and Recall across these datasets indicates that nnUNet may require dataset-specific adjustments or tuning to optimise its effectiveness. This discrepancy calls into question the generability of the model, suggesting that while it can adapt to specific types of data, it may face significant challenges when confronted with other kinds. This pattern underscores the need for a rigorous evaluation and potential refinement of the model to ensure reliable and consistent performance across diverse clinical contexts. Such refinement is crucial for practically deploying the model in real-world medical settings.

The statistical significance columns, denoted by letters, reveal intricate relationships. This suggests that no single model consistently outperforms others across all metrics and datasets. However, specific trends are evident. These include the robust performance of FastSurferCNN (TF) and Dense UNET (TF) in multiple metrics. Furthermore, significance testing based on the paired Wilcoxon signed-rank test underscores the comparative efficacy of these models under different conditions. The complex interplay of performance across different datasets and metrics is essential for an understanding of the generability and robustness of these segmentation models in medical imaging.

## 4 Discussion

In this study, we conducted a thorough examination with the goal of determining the most suitable kernel size for our segmentation models of white matter hyperintensities (WMH). Our findings revealed that a 3×3 kernel size represents an optimal balance between Dice coefficient, 95th percentile Hausdorff distance, and volumetric symmetry. This kernel size offered an optimal trade-off between segmentation accuracy and computational efficiency. It has been established that larger kernels have the capability of capturing a greater spatial context; however, the small area of WMH lesion presents challenges for the extraction of meaningful features with larger kernels. Consequently, the 3x3 kernel size strikes a balance, ensuring precise segmentation without unnecessary computational overhead.

Futhermore, we examine the differential spatial information provided by various MRI modalities and to assess their individual and combined effects on model

performance. The findings indicated that Flair images are particularly effective for lesion segmentation, consistently achieving higher levels of accuracy. This aligns with current literature, which highlights Flair as the gold standard. The use of two modalities resulted in enhanced performance, indicating that combining spatial information from different sources allows for a more comprehensive and detailed representation of anatomical structures. Nevertheless, the combination of the three spatial domains (Flair-T1-T2) yielded improved metrics but also led to higher computational costs. This trade-off should be considered in clinical settings with limited resources. Based on our findings, we recommend using Flair images as the primary modality for lesion segmentation, with Flair combined with another modality such as T1 to enhance accuracy. This combination offers the optimal balance between improved performance and reduced computational costs.

The primary analysis of segmentation performance, utilizing the optimal model from the ablation test with alternating Flair and Flair-T1 inputs, demonstrated that models leveraging Flair as a singular input, particularly the nn-UNet model, achieve effective segmentation overlap and boundary delineation. The addition of the T1 modality generally enhanced performance when used alongside Dense UNET. Incorporating T1 data alongside Flair data has been shown to enhance overall model performance, suggesting that combining data from different spatial domains provides a more comprehensive understanding of anatomical structures. However, it should be noted that the evaluation was based on a specific dataset with a limited number of modalities, which might not capture the full potential or limitations of these models in different contexts. Additionally, the presence of outliers in several metrics indicates occasional performance deviations, requiring further investigation to identify and mitigate these issues. In light of these findings, we recommend the use of multimodal inputs to enhance segmentation accuracy and boundary precision.

As the data lacks considerable demographic diversity, the impact of a more extensive and diverse dataset was investigated, such as the UK Biobank. The utilisation of transfer learning models, such as FastSurferCNN (TF) and Dense UNET (TF), demonstrated superior Dice scores and Recall compared to their non-transfer learning counterparts. This suggests that transfer learning effectively facilitates model generalisation by leveraging pre-learnt features from a psuedo label dataset such a UK Biobank. In particular, FastSurferCNN (TF) exhibited balanced performance across the entirety of the metrics, thus indicating its robustness and generalisability. The enhanced performance of models with transfer learning serves to highlight the potential of this approach to enhance segmentation accuracy and to reduce false negatives, both of which are crucial in clinical settings. Based on the findings of this study, it is recommended that the use of a transfer learning approach be considered for incorporation into segmentation models, particularly where the availability of labelled data is restricted.

At last, we assests the robustness and applicability of the best model from all our analysys across different domain. The results indicate that transfer learning significantly enhances the performance of segmentation models. FastSurferCNN (TF) achieved the highest Dice score and Recall on the ADNI dataset, demonstrating its ability to accurately segment and detect lesions. Similarly, on the MICCAI 2016 dataset, FastSurferCNN (TF) and FastSurferCNN performed exceptionally well in terms of Dice and VS, indicating robust segmentation and volumetric consistency. In the larger MICCAI 2017 dataset, Dense UNET outperformed other models in Dice scores, suggesting its effectiveness in larger datasets. However, the nn-Unet model exhibited high variability, performing poorly on the ADNI dataset but improving significantly on the MICCAI datasets. The findings underscore the importance of transfer learning in enhancing model performance, particularly in datasets with varying characteristics. The variability in nn-Unet's performance across different datasets highlights the challenge of generalizability, since this model rely on the adaptation of the trainning paramaters with the signature reults in a overfitting and also the fewer convolutions in the feature extration influence the lack of meanfull features. Based on the findings, we recommend incorporating transfer learning into segmentation models to enhance their robustness and accuracy. Future research should focus on optimizing transfer learning techniques to reduce computational costs and improve generalizability. It is also important to evaluate models on a wider range of datasets to ensure their reliability across different clinical scenarios. For models like nn-Unet, further investigation into dataset-specific tuning and adjustment strategies is necessary to enhance performance consistency.

## 5 Conclusions

In conclusion, we demonstrated that automated WMH segmentation using deep learning can be achieved when a Flair image is available. Moreover, the performance of state-of-the-art UNet-like architectures is similar regardless of the input modality.

A critical challenge for WMH segmentation is the selection of the training dataset, as there is a lot of anatomical variation across subjects. We demonstrated that all models equally struggled with lesions that are outside the distribution of the training data. Future research should focus on data harmonization between labels and datasets to improve performance in a general population.

Future work should also focus on extending the input flexibility of our tools to broader scenarios, such

as extending the capabilities of the current networks into hetero-modal. Since the segmentation could benefit the deployment of WMH models in a wider range of datasets, as segmentation could be generated even if one of the modalities is missing. Additonally, models that support the hetero-modal could enhance segmentation performance by integration of more datasets without the same modalities. Furthermore, models should explore the separation of the segmentation tasks by size of lesion since the small lesion segmentation sometimes get overtaken by bigger size predictions, instead of relying of the generalization of one model to all sizes of lesions. Inclusion of other lesions could reduce the missclassification of regions closed to the WMHs that mimic their apperace.

Overall, WMH segmentation using deep learning is a feasible task. However, there is still a need for more robust methods that can generalize better across subjects and datasets.

## Acknowledgments

I would like to express my heartfelt gratitude to my supervisors, Santiago and Martin, for their time, attention, and encouragement throughout my master's thesis project. Their guidance and support were invaluable, and I am deeply thankful for the opportunity to learn and work under their mentorship. Additionally, I am grateful to the MAIA program for giving me the chance to pursue this master's degree, experience life in different countries, and meet amazing people. To all the new friends I've made during this two-year journey, thank you for being part of this adventure and for your support. I also extend my appreciation to my friends and people back in Mexico who always offer encouraging words. A mis padres y hermanos, que sin ellos y sin su apoyo incondicional y amor no estaría donde estoy ahora.

## 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., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D.C., Zhang, H., Dragonu, I., Matthews, P.M., Miller, K.L., Smith, S.M., 2018. Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. NeuroImage 166, 400–424. doi:10.1016/j.neuroimage.2017.10.034.

Alzheimer's Disease Neuroimaging Initiative, 2024. Adni. URL: https://adni.loni.usc.edu/.

Breteler, M.M., Stöcker, T., Pracht, E., Brenner, D., Stirnberg, R., 2014. Ic-p-165: Mri in the rhineland study: A novel protocol for population neuroimaging. Alzheimer's & Dementia 10, P92–P92. doi:10.1016/j.jalz.2014.05.172.

Commowick, O., Kain, M., Casey, R., Ameli, R., Ferré, J.C., Kerbrat, A., Tourdias, T., Cervenansky, F., Camarasu-Pop, S., Glatard, T., et al., 2021. Multiple sclerosis lesions segmentation from multiple experts: The miccai 2016 challenge dataset. Neuroimage 244, 118589.

Crimi, A., Bakas, S. (Eds.), 2017. Proceedings of the MICCAI 2017 Brain Lesion Workshop, Springer, Quebec City, Canada.

Debette, S., Markus, H.S., 2010. The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. BMJ 341. doi:10.1136/bmj.c3666.

Estrada, S., Conjeti, S., Ahmad, M., Navab, N., Reuter, M., 2018. Competition vs. concatenation in skip connections of fully convolutional networks, in: Machine Learning in Medical Imaging: 9th International Workshop, MLMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 9, Springer. pp. 214–222.

Fischl, B., 2012. Freesurfer. Neuroimage 62, 774–781. doi:10.1016/j.neuroimage.2012.01.021.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., et al., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355. doi:10.1016/S0896-6273(02)00569-X.

Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., Schulz, U.G., Kuker, W., Battaglini, M., Rothwell, P.M., Jenkinson, M., 2016. Bianca (brain intensity abnormality classification algorithm): A new tool for automated segmentation of white matter hyperintensities. NeuroImage 141, 191–205. doi:10.1016/j.neuroimage.2016.07.018.

Heinen, R., Steenwijk, M.D., Barkhof, F., Biesbroek, J.M., van der Flier, W.M., Kuijf, H.J., Prins, N.D., Vrenken, H., Biessels, G.J., de Bresser, J., et al., 2019. Performance of five automated white matter hyperintensity segmentation methods in a multicenter dataset. Scientific reports 9, 16742. doi:10.1038/s41598-019-53273-x.

Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., Reuter, M., 2020. Fastsurfer - a fast and accurate deep learning based neuroimaging pipeline. NeuroImage 219, 117012. doi:10.1016/j.neuroimage.2020.117012.

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. doi:10.1038/s41592-020-01008-z.

Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y., 2017. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1175–1183. doi:10.1109/CVPRW.2017.156.

Kuijf, H., Biesbroek, M., de Bresser, J., Heinen, R., Chen, C., van der Flier, W., Barkhof, Viergever, M., Biessels, G.J., 2022. Data of the White Matter Hyperintensity (WMH) Segmentation Challenge. URL: https://doi.org/10.34894/AECRSD, doi:10.34894/AECRSD.

Medical Image Computing and Computer Assisted Intervention Society (MICCAI), 2017. Miccai challenge on brain lesion. URL: http://www.miccai.org/challenges-on-brain-lesion. accessed: 2024-05-23.

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.R., Griffanti, L., Douaud, G., Okell, T.W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., Matthews, P.M., Smith, S.M., 2016. Multimodal population brain imaging in the uk biobank prospective epidemiological study. Nature Neuroscience 19, 1523–1536. doi:10.1038/nn.4393.

Park, G., Hong, J., Duffy, B.A., Lee, J.M., Kim, H., 2021. White matter hyperintensities segmentation using the ensemble u-net with multi-scale highlighting foregrounds. NeuroImage 237, 118140. doi:10.1016/j.neuroimage.2021.118140.

Roy, A.G., Conjeti, S., Sheet, D., Katouzian, A., Navab, N., Wachinger, C., 2017. Error corrective boosting for learning fully convolutional networks with limited data, in: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (Eds.), Medical Image Computing and Computer Assisted Intervention, Springer International Publishing, Cham. pp. 231–239.

Safarov, S., Whangbo, T.K., 2021. A-denseunet: Adaptive densely

connected unet for polyp segmentation in colonoscopy images with atrous convolution. Sensors 21, 1441.

Scheltens, P., Barkhof, F., Valk, J., Algra, P., HOOP, R.G.V.D., Nauta, J., Wolters, E.C., 1992. White matter lesions on magnetic resonance imaging in clinically diagnosed alzheimer's disease: evidence for heterogeneity. Brain 115, 735–748. doi:10.1093/brain/115.3.735.

Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., Hoshi, M., Ilg, R., Schmid, V.J., Zimmer, C., Hemmer, B., Mühlau, M., 2012. An automated tool for detection of flair-hyperintense white-matter lesions in multiple sclerosis. NeuroImage 59, 3774–3783. doi:10.1016/j.neuroimage.2011.11.032.

Shiee, N., Bazin, P.L., Cuzzocreo, J., Ye, C., Kishore, B., Carass, A., Calabresi, P., Reich, D., Prince, J., Pham, D., 2014. Reconstruction of the human cerebral cortex robust to white matter lesions: Method and validation. Human Brain Mapping 35. doi:10.1002/hbm.22409.

Shiee, N., Bazin, P.L., Ozturk, A., Reich, D.S., Calabresi, P.A., Pham, D.L., 2010. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. NeuroImage 49, 1524–1535. doi:10.1016/j.neuroimage.2009.09.005.

Smith, J.E., Doe, J., 2018. A review of the miccai 2017 challenge on white matter hyperintensities segmentation. Journal of Medical Imaging 5, 012003. doi:10.1117/1.JMI.5.1.012003.

Steenwijk, M.D., Pouwels, P.J., Daams, M., van Dalen, J.W., Caan, M.W., Richard, E., Barkhof, F., Vrenken, H., 2013. Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (s). NeuroImage: Clinical 3, 462–469. doi:10.1016/j.nicl.2013.10.003.

Stöcker, T., 2016. Big data: The rhineland study, in: Proceedings of the 24th Scientific Meeting of the International Society for Magnetic Resonance in Medicine, Singapore. URL: https://cds.ismrm.org/protected/16MProceedings/PDFfiles/6865.html.

Sundaresan, V., Dinsdale, N.K., Jenkinson, M., Griffanti, L., 2022. Omni-supervised domain adversarial training for white matter hyperintensity segmentation in the uk biobank, in: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1–4. doi:10.1109/ISBI52829.2022.9761540.

Taha, A.A., Hanbury, A., 2015. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. BMC Medical Imaging 15. doi:10.1186/s12880-015-0068-x.

Tran, P., Thoprakarn, U., Gourieux, E., dos Santos, C.L., Cavedo, E., Guizard, N., Cotton, F., Krolak-Salmon, P., Delmaire, C., Heidelberg, D., Pyatigorskaya, N., Ströer, S., Dormont, D., Martini, J.B., Chupin, M., 2022. Automatic segmentation of white matter hyperintensities: validation and comparison with state-of-the-art methods on both multiple sclerosis and elderly subjects. NeuroImage: Clinical 33, 102940. URL: https://www.sciencedirect.com/science/article/pii/S2213158222000055, doi:10.1016/j.nicl.2022.102940.

Trejo-Lopez, J.A., Yachnis, A.T., Prokop, S., 2022. Neuropathology of alzheimer's disease. Neurotherapeutics 19, 173–185. doi:10.1007/s13311-021-01146-y.

Tsuchida, A., Boutinaud, P., Verrecchia, V., Tzourio, C., Debette, S., Joliot, M., 2023. Early detection of white matter hyperintensities using shiva-wmh detector. Preprint on bioRxiv URL: https://doi.org/10.1101/2023.02.03.526961, doi:10.1101/2023.02.03.526961.

Valverde, S., Cabezas, M., Roura, E., González-Villà, S., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Rovira, À., Oliver, A., Lladó, X., 2019. Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach. NeuroImage 186, 445–455. doi:10.1016/j.neuroimage.2018.11.008.

Wardlaw, J.M., Hernández, M.C.V., Muñoz-Maniega, S., 2015. What are white matter hyperintensities made of? Journal of the American Heart Association 4, e001140. doi:10.1161/JAHA.114.001140.

Wilcoxon, F., 1992. Individual comparisons by ranking methods, in: Breakthroughs in statistics: Methodology and distribution. Springer, pp. 196–202. doi:10.1007/978-1-4612-4380-9_16.

Yeung, M., Sala, E., Schönlieb, C.B., Rundo, L., 2022. Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. Computerized Medical Imaging and Graphics 95, 102026. doi:10.1016/j.compmedimag.2021.102026.

## Appendix A Figures

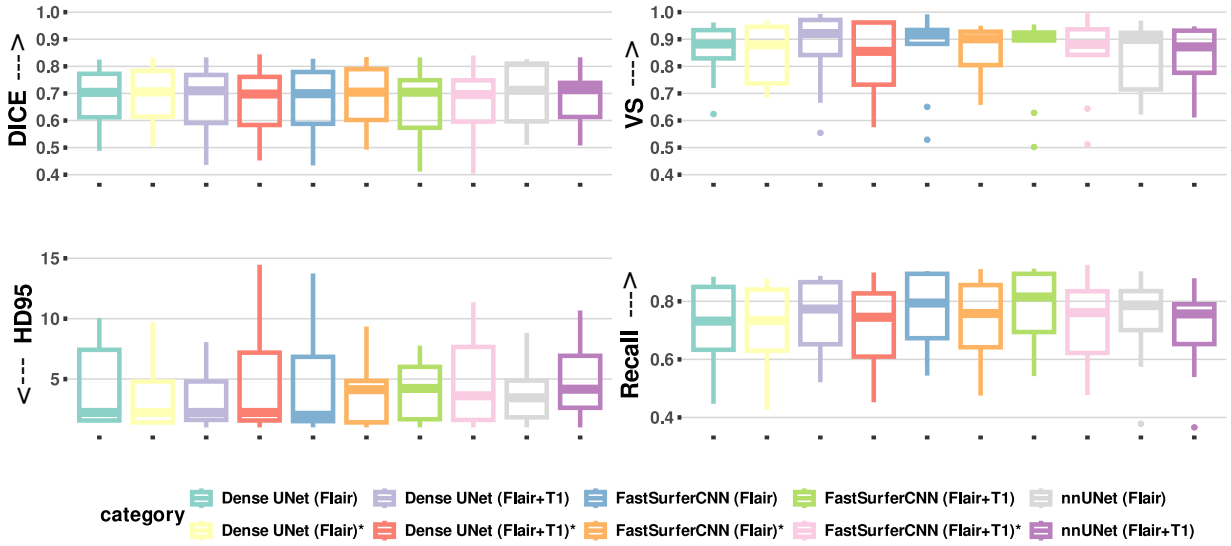



Figure A.10: Segmentation performance comparison on the in-house test-set between the proposed Dense UNET( blue, yellow, illiac and red) and benchmark F-CNNs (orange, dark blue, pink and gren) and nn-UNet (purple and gray) Note: The models mark with a "*" represent the transfer learning models with the base learning from the UK-Biobank

Medical Imaging and Applications

Master Thesis, July 2024

# Few shot probabilistic despeckling in optical coherence tomography

Anita Zhudenova

Supervisors: Néstor Uribe-Patarroyo[1,2], Bhaskara Rao Chintada[1,2]

[a] *Wellman Center for Photomedicine, Massachusetts General Hospital, Boston, MA 02114, USA*
[b] *Harvard Medical School, Boston, MA 02115, USA*

**Abstract**

The capability to image the transparent cells in the inner retina, in particular ganglion cells, has the potential to revolutionize the diagnosis and monitoring of neurodegenerative diseases with ocular presentation, including glaucoma and Alzheimer's disease. Imaging ganglion cells is highly challenging due to their very low intrinsic scattering contrast, which makes them virtually invisible among the strong speckle present in optical coherence tomography (OCT). Currently, speckle reduction using volume averaging of hundreds of OCT volumes is necessary to supress speckle contrast and enable their observation and counting. In this project, we have tested Classical and Machine Learning despeckling methods that uses the observation of a few independent speckle realizations for more accurately estimating despeckled images. This new method will lower imaging requirements for ganglion cell counting by an order of magnitude, from hundreds to tens of OCT volumes.

*Keywords:* Matlab, OCT system, statistics, coherent imaging, maximum likelihood estimation, machine learning, Alternative methods for calculating similarity criteria in TNode OCT despeckling algorithm

## 1. Introduction

Optical Coherence Tomography (OCT) 1 is well- established diagnostic imaging tool that produces a high-resolution cross-sectional image and three- dimensional volumetric measurements of biological tissue and used for the diagnosis of a variety of diseases, such as glaucoma, macular degeneration, early detection of neurodegenerative diseases and macular edema(1). Apart from ophthalmology, OCT is extensively utilized in cardiology, dermatology, and gastroenterology, providing real-time, in situ imaging of tissues. OCT can function as an optical biopsy tool in medical imaging diagnostics, offering detailed images without the need for tissue removal and microscopic analysis, unlike traditional histopathological analysis(7).

OCT is similar to ultrasound imaging, but it uses light instead of sound. All OCT systems use coherent light to measure the echo time delay and the intensity of backscattered or backreflected light from internal microstructures in materials or tissues [see Fig. 2].(7).

Visual interpretation is hindered and the quality and contrast of the image are decreased which is caused by the presence of speckle, a high-contrast, fine-scale granular pattern (3), which is present in OCT images due to its coherent nature (4) [see Fig. 1]. Speckle in OCT is not an additive noise that can be easily removed (19), in fact it carries information about the microstructural properties of the tissue.

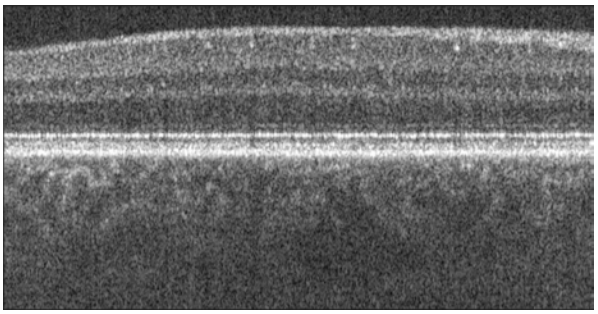


Figure 1: OCT image.

The techniques that suppresses speckle and speckle-like noise while preserving the visibility of fine-scale

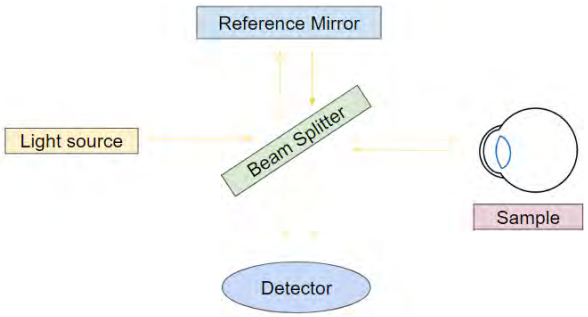



Figure 2: OCT image.

structure have been a research interest of the OCT community for many years (5). The speckle reduction methods aim to expedite the visual interpretation of the image by increasing the quality and making the boundaries between tissues more distinguishable and improving the diagnostic utility in medical applications. Speckle reduction methods are applicable not only for OCT images but also for synthetic aperture radar and and ultrasound imaging. (6).

Speckle significantly impacts the interpretation of the information within an image and makes boundaries between highly scattering structures in tissue difficult to resolve. Speckle occurs in a signal when it is made up of numerous independently phased additive complex components. These components can have both random lengths and random directions (phases) in the complex plane, or they may have known lengths and random directions. When these components are summed, they create a phenomenon known as a "random walk" (3; 4).

### *1.1. Clinical Application*

In clinical diagnosis, OCT is widely used due to its capability to provide real-time, non-invasive, and high-resolution images. It is used for the diagnostic of neurodegenerative diseases because it can capture the changes in retina in Alzheimer's disease (AD) and mild cognitive impairment (MCI) patients, such as ganglion cell degeneration, thinning of the retinal nerve fiber layer (RNFL), optic nerve axonal loss, and decreased perfusion and vessel density (24; 25). OCT can quantitatively assess the thickness of the Ganglion Cell Layer (GCL) and RNFL.

Geographic atrophy (GA) is a severe form of Age-related Macular Degeneration (AMD), marked by chronic progressive degeneration of the macula. The condition can be characterized by atrophic lesions that initially develop in the outer retina and gradually extend to cover the macula and the fovea, the central area of the macula, ultimately leading to permanent vision loss. OCT detects these lesions through the loss of retinal layers, which can be observed in cross-sectional and/or en-face images. This allows for a detailed assessment of the features of GA lesions (26; 27).

Intravascular OCT (IV-OCT) uses a catheter-based OCT system to image the coronary arteries, providing detailed information about plaque composition, fibrous cap thickness, and stent apposition (22).

## 2. State of the art

Speckle-suppression methods can be classified into two main categories: hardware-based and post-processing methods. In the first approaches, the systems are modified to produce uncorrelated speckle patterns within or between B-scans, a series of longitudinal scans (A-scans) (1; 6). If B-scans are acquired closely and rapidly, they can be translated into a volumetric image or C-scan [see Fig. 3].

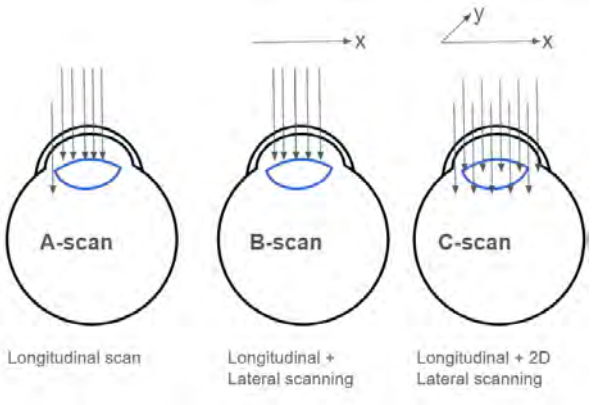



Figure 3: A-scan, B-scan and C-scans of the eye.

Conversely, the hardware methods can be divided into subcategories:

1. Angular compounding: images are acquired at different backscattering angles and combined to improve the Signal-To-Noise ratio. Requires precise alignment and calibration of the imaging system to accurately combine images, as a result, the complexity and acquisition time increases.(18; 20).
2. Spatial compounding: the absolute magnitudes of signals obtained from the same or slightly shifted sample volumes are averaged to create a new signal with reduced speckle noise (4).
3. Frequency compounding: benefits from the reduced correlation between two speckled images acquired at different optical frequency band (4; 9).
4. Polarization diversity: enhance image quality by exploiting the polarization properties of light and can be achieved by using unpolarized light to illuminate the sample and interfering the backscattered light with an unpolarized reference beam(4)

While those methods that reduce the speckle contrast 5, a significant amount of information is lost in each implementation: in spatial compounding, the neighboring sample locations are joined into a single pixel and the

SNR, the speckle-to-noise ratio, gain can be reduced by any correlation among the signals; in frequency compounding, the raw data contains information that belongs to a higher axial resolution (4; 6).

Current speckle suppression post-processing algorithms can be categorized into following families:

1. Transformation based: wavelet or curvelet transforms to represent images in a different domain.
2. Sparse representation: special scanning technique where few B-scans acquired with high nominal SNR are used to improve the quality of low SNR B-scans (21) (6)
3. Spatial domain: primarily involve non-local methods, in which each pixel is assessed within an extensive neighborhood that leverages the inherent redundancy of natural images (6; 10–13; 15; 16)

In transform-based methods, it is assumed that the coefficients representing the image are distinguishable in the curvelet or wavelet domain. The is difficult to choose a correct threshold for the speckle coefficients, the wrong selection causes the point-like structures to disappear in the filtered image and to preserve the contrast tissue layers (6; 17; 21; 23).

Other post-processing algorithms include using Deep Learning for despeckling OCT images: they can effectively remove speckles and preserve the important structural details. However, obtaining ground truth for supervised models is challenging and there is a risk for over-smoothing since some of the models remove the fine structures. As well as that the models are not generalized for all OCT systems, since the parameters for every system is different.

Important to note, that all post-processing methods require prior averaging of multiple speckle realizations, which increases the computational time of this type of approach. The goal of the project is to find a new way of leveraging information from multiple speckle realizations inside of TNode (Tomographic Non-local-means despeckling), a powerful volumetric non-local means despeckling algorithm.

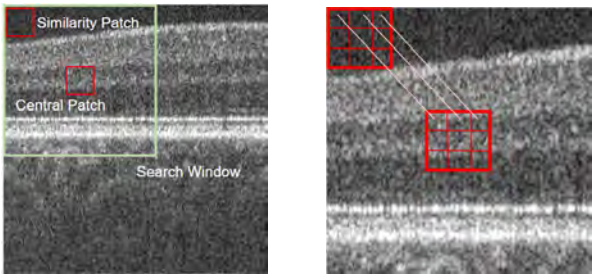
## 3. Material and methods

### *3.1. Conventional TNode*

TNode algorithm build upon the polarimetric and interferometric non-local framework build by Deledalle et al (6; 10–12) for SAR images. The fundamental concept is to identify small volumetric patches within the tomogram that represent various speckle realizations of the same underlying structure, such as tiny blood vessels. These patches are used for non-local, incoherent averaging. This approach ensures that only uncorrelated speckle realizations of the same object are combined, preserving resolution and reducing speckle contrast.

Using speckle statistics, we can determine the speckle-free intensity from a general context that accommodates both single and multiple speckle realization tomograms. This versatile framework enables the application of TNode in-vivo for single-look frame analysis and can be easily adapted to scenarios with multiple speckle realizations (6), by taking the average of the images prior to processing.

By using 3D similarity windows to retrieve the weights from the volumetric patch-similarity, TNode 4a exploits the available volumetric information in OCT and ensures that one- and two- dimensional structures, that in traditional OCT imaging appear one-dimensional due to resolution limitations, are correctly characterized and preserved upon despeckling(6).




(a) Search window defined within a tomogram.

(b) Pixel by pixel comparison between similarity and central patches within a search window.

Figure 4: Similarity value calculation between pixel of interest x and x' in TNode.

#### *3.1.1. Non-local denoising and speckle suppression*

Non-local methods denoise the image by performing a weighted intensity average in the set of regions, called patches that are ideally anywhere in the image (6; 11; 12; 16). Non-local methods are based on the probability that the two given patches are the different realizations of noise and have the same underlying object, and in the case of non-local despeckling, it means that pixels within those patches are described by a single probability distribution with a common parameter $\theta$, . To define the probability of a given pixel to share the parameter $\theta$ with other pixel x' within the neighborhood or search window, the intensity of pixel x is compared with the intensity of the pixel within the set of search windows.

In TNode, the group of pixels $\mathbf{x}$ centered around a pixel of interest $x$ is compared with the pixels $\mathbf{x}'$ inside of a patch centered around $x'$. It is important to choose the optimal size of the patch that will have enough independent speckles to obtain a robust similarity metric but not too large to reduce the effectiveness of the despeckling method.

The parameter $\theta$ in practice is unknown, instead, it is

replaced by the generalized likelihood ratio (GRL) $\mathscr{L}G$.

$$\mathcal{L}_G\left[\mathbf{I}(x),\mathbf{I}(x')\right]=\left(\frac{\mathbf{I}(x)\mathbf{I}(x')}{\left(\frac{1}{2}\left[\mathbf{I}(x)+\mathbf{I}(x')\right]\right)^2}\right)^L. \quad (1)$$

In case of identical patches, the $\mathscr{L}G = 1$. The GRL values are used to map the similarities between patches into weights. The project aims to find an alternative way to calculate those values using Classical and Machine Learning models.

After calculating GRL values, to obtain the weights, first, the probabilities are compounded inside of a patch into a log-probability $\Delta(\vec{x},\vec{x}')$.

$$\Delta(\vec{x},\vec{x}')=\sum_{\vec{\tau}\in\vec{p}}\log\left(\mathcal{L}_G\left[I(\vec{x}+\vec{\tau}),I(\vec{x}'+\vec{\tau})\right]\right), \quad (2)$$

where $(\vec{x},\vec{x}')$ are the vectors indicating the location of the pixel in the volume and $\vec{\tau}$ is a 3D shift vector.

After computing the sum of log probabilities using 2, weights are obtained by using the following equation:

$$w(\vec{x},\vec{x}')=\exp\left[\frac{\Delta(\vec{x},\vec{x}')}{h}\right], \quad (3)$$

where $h > 0$ and controls the overall distribution of weight values: large **h** increases the probability ratio for each pixel by equalizing the contribution of weights. Important to note that TNode is non local since the weights are affected only by the patch-similarity without taking into consideration the relative distance between patches (6).

The code is available at `https://doi.org/10.6084/m9.figshare.6089861.v1`.

### 3.2. Speckle metrics

There are two metrics to quantify the deleterious effect of speckle in the OCT images. The first one is the Signal-to-speckle ratio (SSR), defined as $s$, an inverse of the speckle contrast:

$$s=\frac{1}{c}=\frac{I}{\theta} \quad (4)$$

where $\bar{I}$ represents the is the speckle-free object intensity, $C$ is the unity speckle contrast in tomograms with a single realization of speckle, and $\theta$ the speckle standard deviation.

Signal-to-noise ratio (SNR), defined as $sn$ ,used to assess the image quality (8) and can be defined as:

$$\mathrm{sn}=\frac{I_s}{I_n} \quad (5)$$

where $I_s$ is the OCT noise-free signal intensity and $I_n$ the noise intensity, which can be all sources of noise (e.g. shot, thermal, digitization, excess photon, etc.) in the OCT system (4; 6; 28; 29).

### 3.3. Data Simulation

The proposed model was trained on simulated tomogram images. The data simulation method can generate single and multiple coherent speckle realizations with different tomogram sizes and numbers of speckle realizations. To ensure the model's generalization, the code utilizes random scale factors for inclusions (scaleInclusion) and the background (scaleBackground). These scales span specified ranges, emulating different signal intensities measured in dB.

A low-pass filtering applied to achieve the spatial correlation and an explicit seed is provided for the random number generator to ensure the consistency of the result across multiple runs.

GRL calculated for the average of 350 and 150 speckle realizations, because averaging more speckle realizations generally leads to better despeckling result were used as a target label5 for training the models that will estimate the similarity value between two patches. Single pixel pairs extracted inside of a search window from 100 simulated tomograms were used as independent variables.

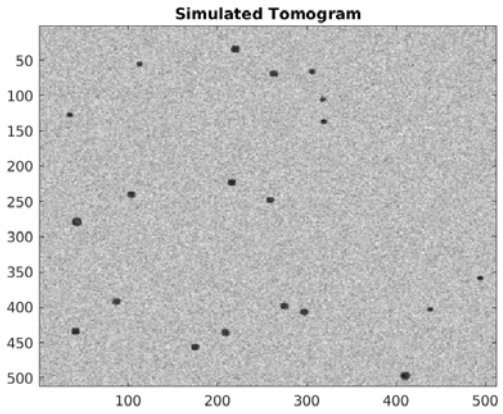



Figure 5: Average of 350 speckle realizations.

### 3.4. Objective

To suppress the speckle contrast it is required to average the hundreds of OCT volumes for speckle reduction. The Conventional TNode processes multiple speckle realizations by first obtaining the mean image. The goal of this project is to determine if there is an alternative way of leveraging the speckle realizations to preserve more information. Two methods were tested on simulated tomograms: the classical approach using conventional TNode and using Neural Networks as alternative method of computing 3.

Three ML methods were tested as an alternative way of calculating the equation 3: 1D Convolutional Neural Network, Siamese Neural Network, and Long short-term memory (LSTM) is a type of Recurrent Neural

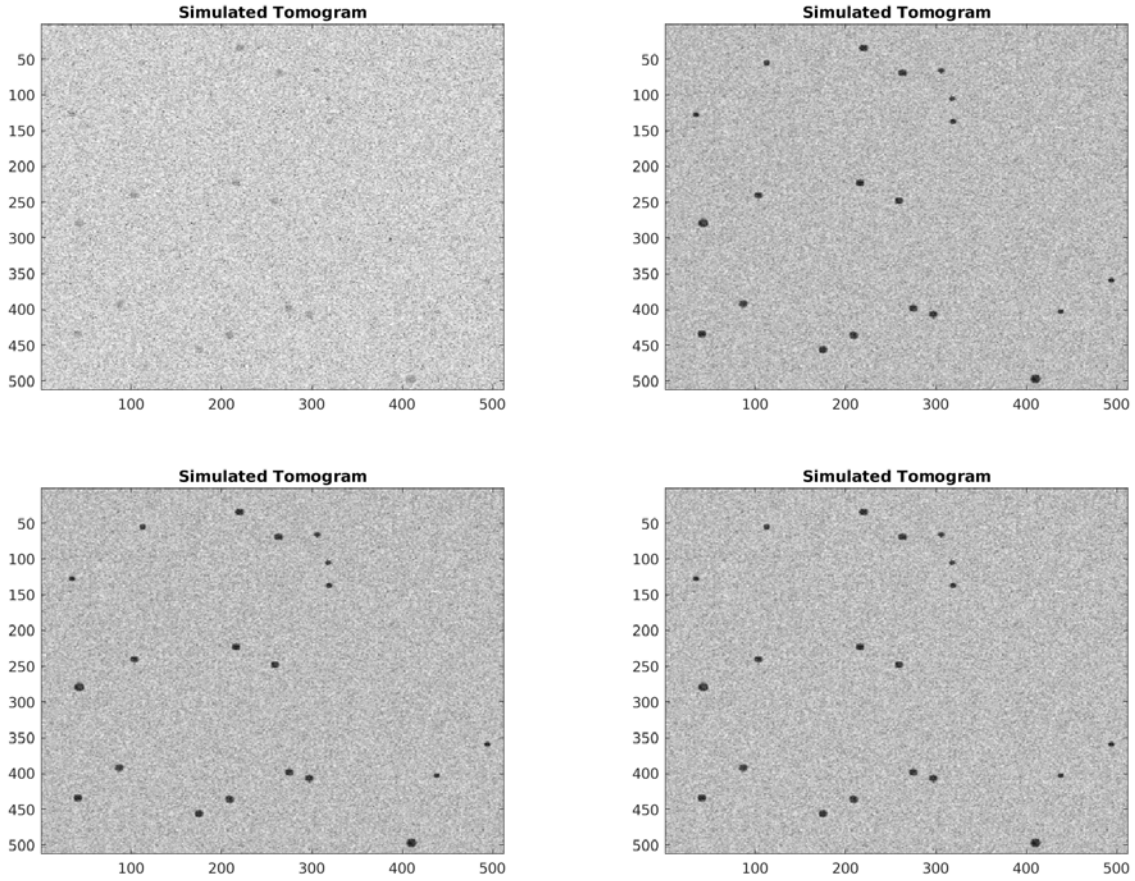



Figure 6: Simulated tomograms showing different numbers of speckle realizations: (a) One speckle realization, (b) Average of ten speckle realizations, (c) Average of 100 speckle realizations, (d) Average of 1000 speckle realizations

Network. The input for the models were 2 vectors: single pixels across 10 speckle realizations extracted from 2 compared patches.

Deep Learning method was implemented in DL-TNode-3D algorithm (2), and is based on a conditional generative adversarial network (cGAN) that leverages the volumetric nature, effectively preserving tissue structures. TNode is very powerful algorithm for OCT despeckling, but it is computationally expensive, incorporation of a deep learning method allows to get speckle suppressed volume two times faster. This method takes a search window with size [2 x 8 + 1] as the generator and TNode despeckled partial volumes as the discriminator.

### *3.4.1. Classical Approach*

As mentioned earlier, conventional TNode processed the average of multiple speckle realizations. Before trying Neural Networks, different leveraging methods were tested: instead of averaging all speckle realizations, GRL was calculated across all speckle realizations creating a similarity matrix and then only the realizations with the highest confidence were averaged in order to estimate the real GRL.

Firstly, the GRL similarity matrix was calculated of a single pixel with itself, between two similar pixel pairs and two dissimilar pixel pairs 11. That was done to analyze the distribution of GRL values inside of a larger structure, search window, inside inclusion, and background.

An unweighted similarity matrix for the whole search window centered around the tested pixel in the inclusion 9 and in the background 10 was obtained by using equation (4), and followed by equation (5) to get the weighted matrix. Those two matrices were compared with the conventional TNode weights, that were obtained by processing the mean of the 10 speckle realizations.

The approach of calculating the GRL values for each speckle realization separately did not produce a better estimation of the real GRL, as can be seen in Fig 7, where despeckling result of prior average of 10 speckle realizations and the result of the proposed method produced the same result. Which led to incorporation of Neural Networks as an alternative for equation 3.

**Algorithm 1** Tomogram Simulation

**Input:** *tomSize*, *numSpeckleRealizations*, *randomSeed*
**Output:** *tomIntLPMultilook*, *groundtruth*, *logLim*
**if** *randomSeed* ≠ empty **then**
    Initialize random generator with *randomSeed*
**end if**
Define noise scale bounds for inclusion and background
Calculate *scaleInclusion* and *scaleBackground* using random values within defined bounds
Initialize parameters for circular masks
Generate non-overlapping circular masks
Define *groundtruth* matrix based on masks and scale values
Initialize *tomIntLPMultilook* matrix
**for** each speckle realization **do**
    Generate complex Gaussian noise
    Scale noise with *groundtruth* to simulate signal
    Apply Fourier transform and a Hanning filter
    Inverse Fourier transform to get low-pass filtered result
    Compute magnitude squared to form tomogram for current realization
**end for**

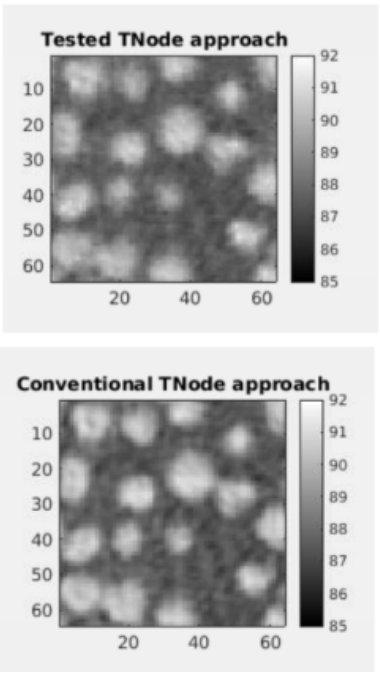
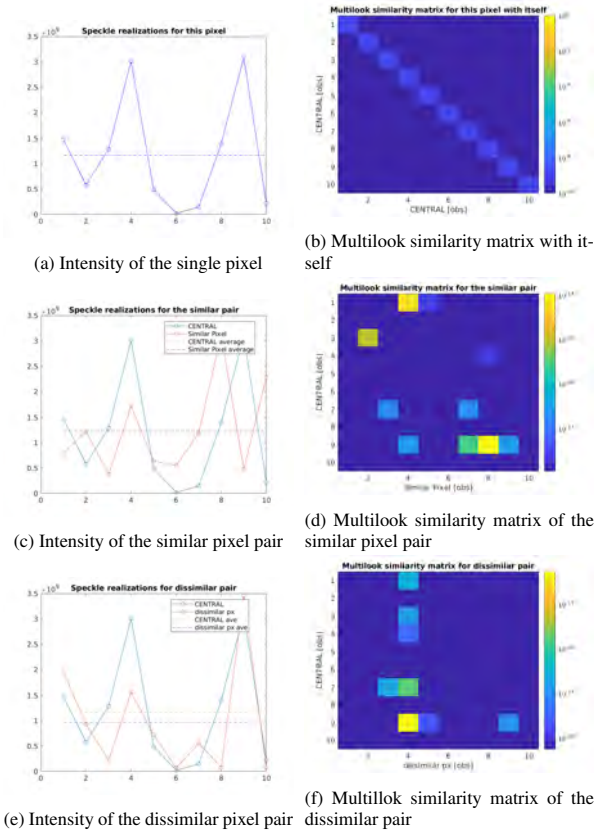



Figure 7: Tested approach vs Conventional TNode




(a) Intensity of the single pixel
(b) Multilook similarity matrix with itself
(c) Intensity of the similar pixel pair
(d) Multilook similarity matrix of the similar pixel pair
(e) Intensity of the dissimilar pixel pair
(f) Multillok similarity matrix of the dissimilar pair

Figure 11: Single pixel pair across 10 speckle realizations

### *3.5. Machine Learning Approach*

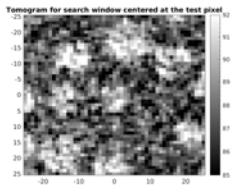
The input for the NN was a time series with the size of [1, 20] and the ground truth was the GRL values calculated by conventional TNode for the averaged 350 speckle realizations 5.

Siamese Neural Networks work by learning a similarity function that compares the feature vectors of two input samples to determine if they are similar or dissimilar. It consists of two or more identical twin networks (sub-networks) with the same architecture, parameters, and weights. Each sub-network extracts relevant features from its input sample through convolutional and pooling layers. The final layer of each sub-network produces a compact feature vector representation of the input. Similarity comparison is done by using a similarity metric, such as Euclidean distance or cosine similarity. This comparison produces a similarity score indicating how similar or dissimilar the two input samples are.

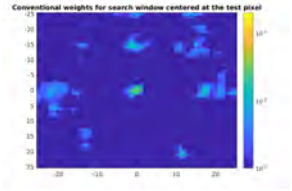
In principle, since the GRL values are basically the estimation of how similar are two patches, Siamese networks should have demonstrated the best result, since this method was created to estimate the similarity between two inputs, for example for one-shot image recognition, proposed by *Koch* (14).

In the case of Neural Networks, for the proposed method One-dimensional convolutional neural

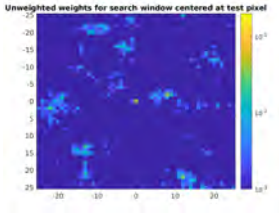

Figure 8: 1D CNN



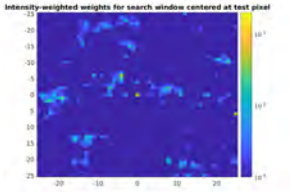

(a) Tomogram for search window centered at the test pixel




(b) Conventional weights for the search window around the tested pixel in the inclusion

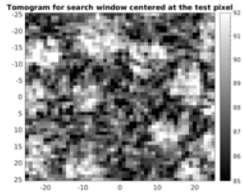



(c) Unweighted weights for the search window around the tested pixel in the inclusion

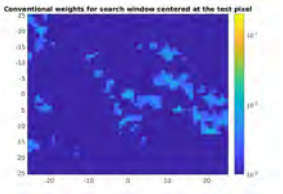



(d) Intensity-weighted weights for the search window around the tested pixel in the inclusion

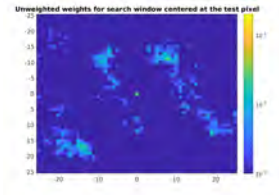
Figure 9: Visual representation of different similarity matrices for the search window centered at the pixel in the inclusion



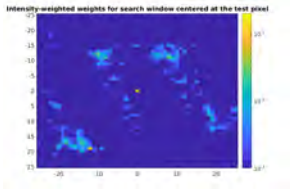

(a) Tomogram for search window centered at the test pixel




(b) Conventional weights for the search window around the tested pixel in the background

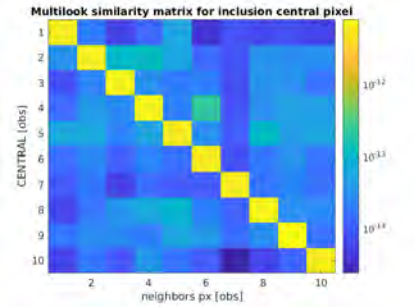



(c) Unweighted weights for the search window around the tested pixel in the background

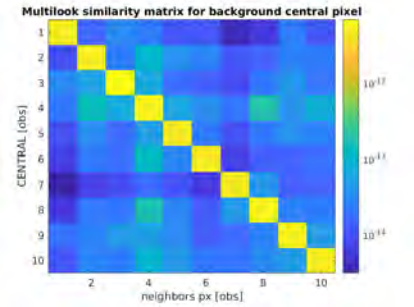



(d) Intensity-weighted weights for the search window around the tested pixel in the background

Figure 10: Visual representation of different similarity matrices for the search window centered at the pixel in the background




(a) Similarity matrix of the central pixel in the inclusion




(b) Similarity matrix of the central pixel in the background

Figure 12: Similarity matrix of 10 speckle realizations

networks (1D CNNs) were used since they are well-suited for time series data and have shown excellent performance in various time series classification and forecasting tasks (31). The model in our case processes the input vectors as two time-series with 10 time steps.

LSTMs are a powerful approach for processing multi-step time series due to their ability to learn long-term dependencies in sequential data.

Table 1: 1D CNN final architecture

| Layer (type) | Output Shape | Param # |
|---|---|---|
| Conv1D | (None, 8, 8) | 56 |
| Conv1D | (None, 6, 12) | 300 |
| Conv1D | (None, 4, 24) | 888 |
| Conv1D | (None, 2, 48) | 3504 |
| MaxPooling1D | (None, 1, 48) | 0 |
| Flatten | (None, 48) | 0 |
| Dense | (None, 15) | 735 |
| Dense | (None, 10) | 160 |
| Dense | (None, 5) | 55 |
| Dense | (None, 1) | 6 |

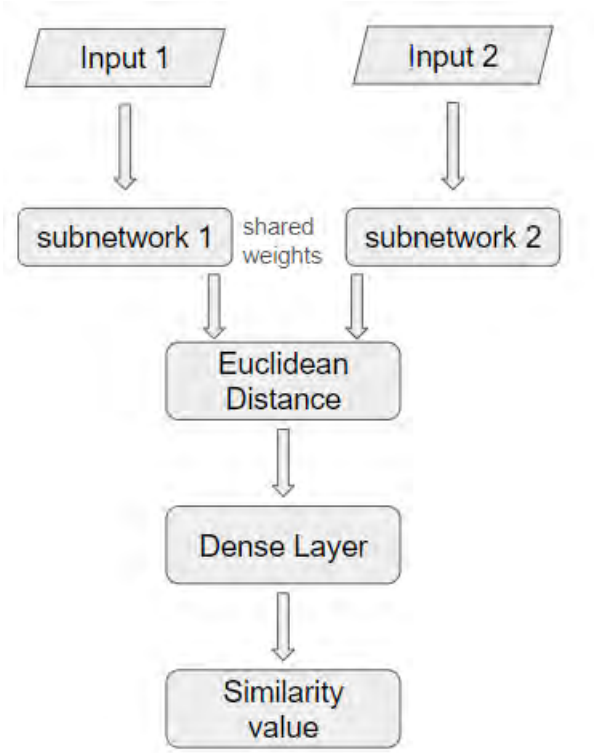



Figure 13: Architecture of Siamese Neural Networks with one input pair(30).

| Layer (type) | Output Shape | Param # |
|---|---|---|
| Bidirectional (Bidirectional) | (None, 10, 16) | 704 |
| Bidirectional (Bidirectional) | (None, 32) | 4224 |
| Dense (Dense) | (None, 15) | 495 |
| Dense (Dense) | (None, 10) | 160 |
| Dense (Dense) | (None, 5) | 55 |
| Dense (Dense) | (None, 1) | 6 |

Table 2: LSTM final architecture

## 4. Results

The final architecture of Siamese Neural Networks is 21 and 15 for the sub-networks, the 1D CNN best architecture is 1 and 14, with 4 1D layers, MaxPooling layer, and 4 fully connected layers. The complexity of the model was gradually increased since lesser layers could not learn the pattern in the data.

Since the GRL values were calculated by processing the average of 350 speckle realizations, to compare the results of NN with the conventional TNode, GRL values calculated just on 10 speckle realizations were used.

The models with batch sizes, the number of samples that are passed to the network, of 32 and 8 perform the best, achieving the highest $R^2$ scores and the lowest RMSE, MSE, and MAE on both validation and test datasets. The model with a batch size of 16 performs the worst across all metrics, indicating it is suboptimal. Batch sizes 32 and 8 show consistent performance between validation and test sets, indicating good generalization. The model with a batch size of 32 achieves the lowest RMSE and MAE, closely followed by batch size 8, suggesting these batch sizes provide more precise predictions with lower error margins.

For Siamese Neural Networks, batch size of 8 shows the best performance, with the highest $R^2$ score and the lowest RMSE, MSE, and MAE. As the batch size increases, there is a slight degradation in performance, indicated by lower $R^2$ scores and higher RMSE, MSE, and MAE values 18. In general, smaller batch sizes (8 and 16) tend to perform better across all metrics compared to larger batch sizes (32 and 128). This suggests that smaller batch sizes allow the model to learn more detailed and nuanced features from the data, leading to better performance. The $R^2$ score is highest for the batch size of 8 (0.3932), indicating that this batch size explains the most variance in the data. The $R^2$ score decreases as the batch size increases, with the lowest $R^2$ score for batch size 128 (0.2965). As the batch size increases, there is a slight degradation in performance, indicated by lower $R^2$ scores and higher RMSE, MSE, and MAE values.

In the case of LSTM, The model begins with two bidirectional LSTM layers. The first layer processes the input sequence in both forward and backward directions, returning the full sequence. The second layer also processes the sequence in both directions but outputs only the final state. These are followed by a series of dense layers with ReLU activation functions, which introduce non-linearity and progressively reduce the data's dimensionality.

The model with LSTM with batch size 32 out-

| Batch Size | 8 | 16 | 32 | 128 |
|---|---|---|---|---|
| $R^2$ Score | 0.3932 | 0.1688 | 0.3594 | 0.2965 |
| RMSE | 0.1626 | 0.1688 | 0.1663 | 0.1742 |
| MSE | 0.0264 | 0.0285 | 0.0276 | 0.0304 |
| MAE | 0.1281 | 0.1333 | 0.1311 | 0.1373 |

Figure 18: Performance metrics for Siamese Network testing different batch sizes. Smaller batch sizes (8 and 32) seem to provide better prediction accuracy and model performance, with batch size 8 being the best among the ones tested.

For three tested models, adding more layers (pooling, normalization and dropout layers) was ineffective and caused the models' overfitting.

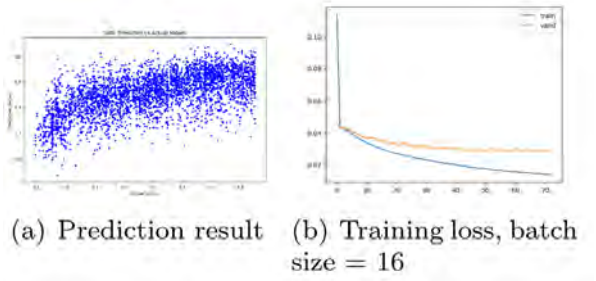
Table 3: Performance metrics for LSTM with different batch sizes, with the batch size= 32 providing the best prediction accuracy and model performance among the tested batch sizes.

| | 8 | 32 | 128 |
|---|---|---|---|
| RMSE | 0.1742 | 0.1688 | 0.1742 |
| MSE | 0.0304 | 0.0285 | 0.0304 |
| MAE | 0.1373 | 0.1333 | 0.1373 |
| $R^2$ | 0.2965 | 0.3393 | 0.2965 |

Figure 14: 1D CNN

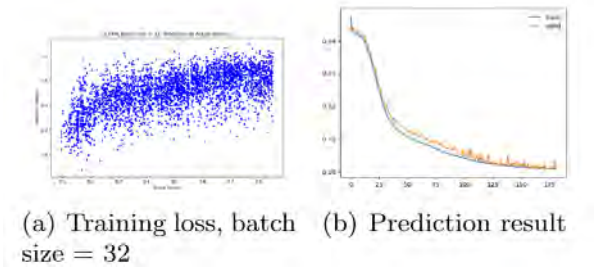
| Layer (type) | Output Shape | Param # | Connected to |
|---|---|---|---|
| InputLayer | [(None, 10, 1)] | 0 | [] |
| InputLayer | [(None, 10, 1)] | 0 | [] |
| Sequential | (None, 64) | 19252 | ['input_3[0][0]', 'input_4[0][0]'] |
| Lambda | (None, 1) | 0 | ['sequential_1[0][0]', 'sequential_1[1][0]'] |
| Dense | (None, 1) | 2 | ['lambda[0][0]'] |

Figure 15: SiameseNN final architecture



(a) Prediction result (b) Training loss, batch size = 16

Figure 16: SNN with batch size = 16. There is a divergence between training and validation loss, prediction result plot shows considerable variance in the prediction.



(a) Training loss, batch size = 32 (b) Prediction result

Figure 17: LSTM with batch size = 32. There is some variability in the predictions, but the training and validation losses are closely aligned, which possibly indicates good generalization.

Table 1: 1D CNN: Validation and Test Data Metrics for Different Batch Sizes

| Batch Size | Data Type | $R^2$ Score | RMSE | MSE | MAE |
|---|---|---|---|---|---|
| 32 | Validation | 0.985652 | 0.025028 | 0.000626 | 0.015350 |
| 32 | Test | 0.982047 | 0.027843 | 0.000775 | 0.015885 |
| 64 | Validation | 0.979181 | 0.030008 | 0.000900 | 0.018209 |
| 64 | Test | 0.979954 | 0.029702 | 0.000882 | 0.018768 |
| 16 | Validation | 0.935595 | 0.052314 | 0.002737 | 0.034763 |
| 16 | Test | 0.935525 | 0.053084 | 0.002818 | 0.035384 |
| 8 | Validation | 0.986911 | 0.023878 | 0.000570 | 0.016293 |
| 8 | Test | 0.986928 | 0.023839 | 0.000568 | 0.016092 |

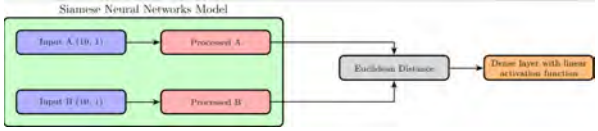



Figure 21: Final SiameseNN architecture.

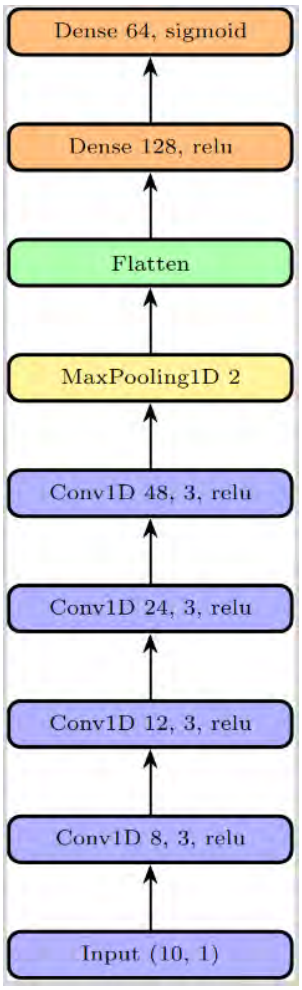
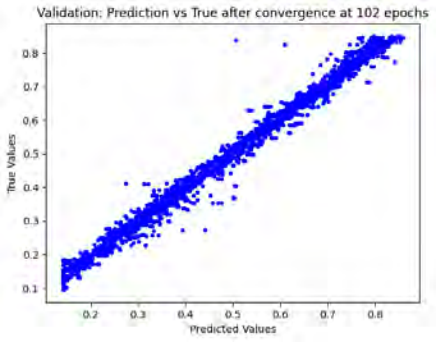


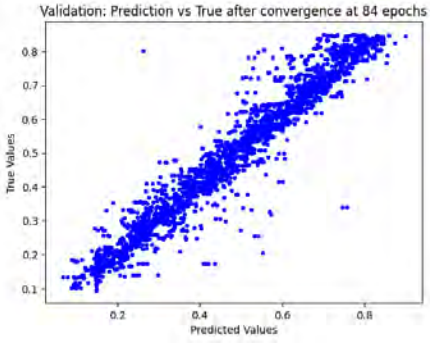

Figure 22: Architecture of SiameseNN sub-networks.

## 5. Discussion

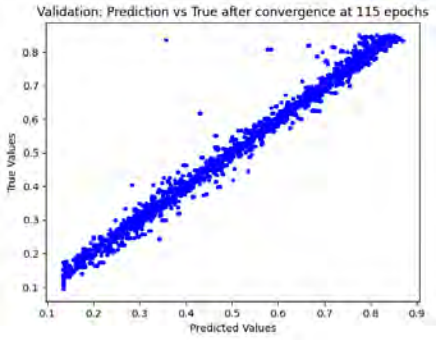
The next step for this project is to process whole volumes and test the model on the real OCT images. More tests should be done for the Siamese Neural Networks with the base models that will take into consideration the independent nature of the speckle realizations.



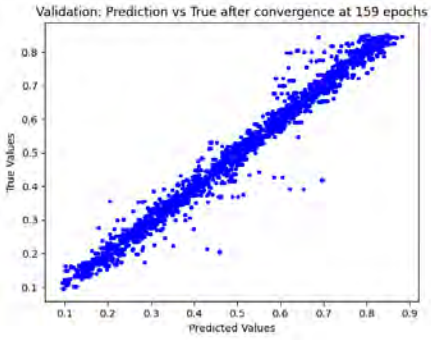

(a) Batch size = 8.




(b) Batch size = 16.




(c) Batch size = 32.




(d) Batch size = 64.

Figure 20: 1D CNN with different batch sizes. There are slight variations in prediction accuracy, the larger batch sizes (32 and 64) tend to produce more accurate predictions with fewer deviations.

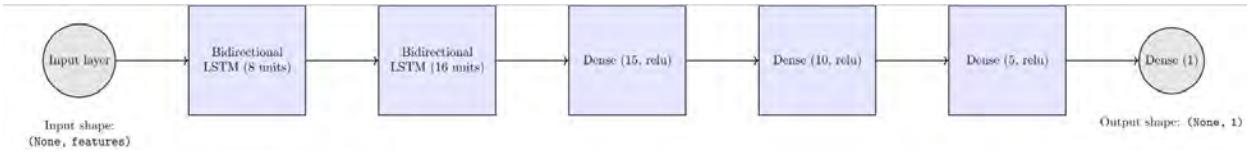



Figure 23: Final LSTM architecture.

Although 1D CNN showed better result compared to other two models, since it takes as an input two features/vectors with 10 time steps, the feature extraction process remains unclear: two vectors are concatenated and then the features are extracted, or the features extracted independently from each vector. This is an important aspect since every pixel in the vector are uncorrelated and should be processed as such.

## 6. Conclusions

The 1D CNN showed the best result in terms of accuracy of the prediction and the speed with best $R^2$ values equals 0.9869. While Siamese Neural Networks seems to be more suitable for this task, it struggled to capture the pattern in the data with $R^2 = 0.39$ which indicates there is a substantial amount of variance not explained by the model. The future work includes implementing SiameseNN with different models that will take into consideration the independent nature of the speckle realizations and will be input permutation invariant.

The implementation of 1D CNN, SiameseNN and LSTM can be be found here

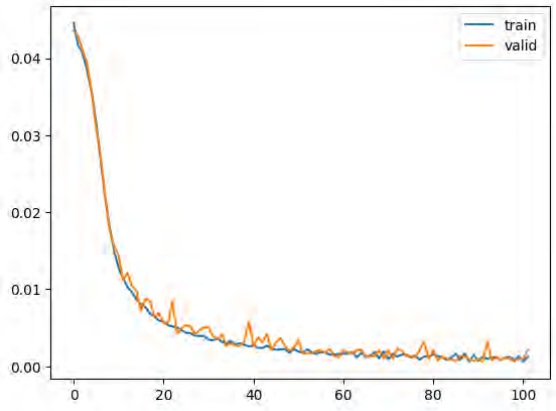
## Acknowledgments

I would like to thank my supervisors at Massachusetts General Hospital for providing me with support through this project and my supervisors at the University of Girona for their support during my Master's program.

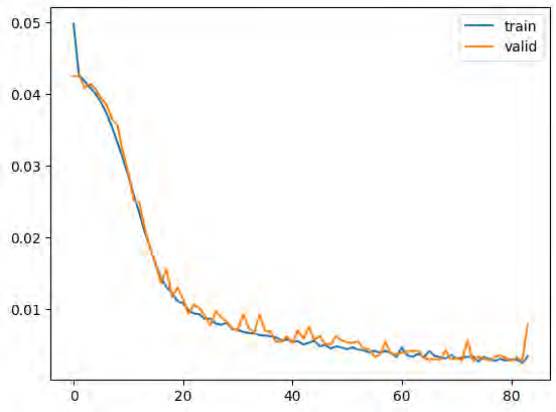
## References

[1] Huang D, Swanson EA, Lin CP, Schuman JS, Stinson WG, Chang W, Hee MR, Flotte T, Gregory K, Puliafito CA, et al. Optical coherence tomography. Science. 1991 Nov 22;254(5035):1178-81. doi: 10.1126/science.1957169. PMID: 1957169; PMCID: PMC4638169.

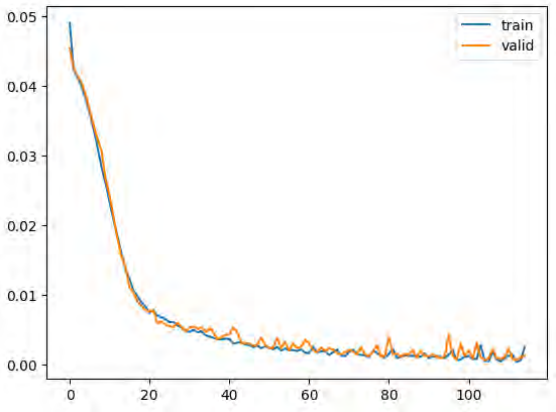
[2] Chintada BR, Ruiz-Lopera S, Restrepo R, Bouma BE, Villiger M, Uribe-Patarroyo N. Probabilistic volumetric speckle suppression in OCT using deep learning. ArXiv [Preprint]. 2023 Dec 7:arXiv:2312.04460v1. PMID: 38106457; PMCID: PMC10723542.




(a) Batch size = 8.




(b) Batch size = 16.




(c) Batch size = 32.

Figure 24: Training and validation loss over epochs with different batch sizes. The performance is consistent across batch sizes, with minimal difference in the final loss values.

[3] J. W. Goodman, *Speckle Phenomena in Optics: Theory and Applications*, Roberts and Company Publishers, 2007.

[4] J. M. Schmitt, S. Xiang, and K. M. Yung, *Speckle in Optical Coherence Tomography*, Journal of Biomedical Optics, vol. 4, pp. 95-105, 1999.

[5] N. Iftimia, B. E. Bouma, G. J. Tearney, "Speckle reduction in optical coherence tomography by 'path length encoded' angular compounding," *Journal of Biomedical Optics*, vol. 8, no. 2, pp. 260-263, Apr. 2003. doi: 10.1117/1.1559060. PMID: 12683852.

[6] C. Cuartas-Vélez, R. Restrepo, B. E. Bouma, N. Uribe-Patarroyo, "Volumetric non-local-means based speckle reduction for optical coherence tomography," *Biomedical Optics Express*, vol. 9, no. 7, pp. 3354-3372, Jun. 2018. doi: 10.1364/BOE.9.003354. PMID: 29984102; PMCID: PMC6033569.

[7] Fujimoto JG, Pitris C, Boppart SA, Brezinski ME. Optical coherence tomography: an emerging technology for biomedical imaging and optical biopsy. Neoplasia. 2000 Jan-Apr;2(1-2):9-25. doi: 10.1038/sj.neo.7900071. PMID: 10933065; PMCID: PMC1531864.

[8] Frosz, Michael, Juhl, Michael, and Lang, Morten. Optical Coherence Tomography: System Design and Noise Analysis. 2001.

[9] Pircher M, Gotzinger E, Leitgeb R, Fercher AF, Hitzenberger CK. Speckle reduction in optical coherence tomography by frequency compounding. *J Biomed Opt*. 2003 Jul;8(3):565-9. doi: 10.1117/1.1578087. PMID: 12880365.

[10] C. A. Deledalle, L. Denis, and F. Tupin, "Iterative Weighted Maximum Likelihood Denoising with Probabilistic Patch-Based Weights," *IEEE Transactions on Image Processing*, vol. 18, no. 12, pp. 2661–2672, Dec. 2009. DOI: 10.1109/TIP.2009.2029593. Epub 2009 Aug 7. PMID: 19666338.

[11] C.-A. Deledalle, L. Denis, F. Tupin, A. Reigber, and M. Jäger, "NL-SAR: A Unified Nonlocal Framework for Resolution-Preserving (Pol)(In)SAR Denoising," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 4, pp. 2021–2038, 2015. DOI: 10.1109/TGRS.2014.2352555.

[12] C.-A. Deledalle, L. Denis, and F. Tupin, "How to Compare Noisy Patches? Patch Similarity Beyond Gaussian Noise," *International Journal of Computer Vision*, vol. 99, pp. 86–102, 2012. Publisher: Springer.

[13] H. Yu, J. Gao, and A. Li, "Probability-Based Non-Local Means Filter for Speckle Noise Suppression in Optical Coherence Tomography Images," *Optics Letters*, vol. 41, no. 5, pp. 994–997, Mar. 2016. DOI: 10.1364/OL.41.000994. PMID: 26974099.

[14] Gregory R. Koch. Siamese Neural Networks for One-Shot Image Recognition. 2015. `https://api.semanticscholar.org/CorpusID:13874643`.

[15] Y. Gu and X. Zhang, "Spiking Cortical Model Based Non-Local Means Method for Despeckling Multiframe Optical Coherence Tomography Data," *Laser Physics Letters*, vol. 14, no. 5, pp. 056201, Apr. 2017. DOI: 10.1088/1612-202X/aa6acf.

[16] A. Buades, B. Coll, and J. M. Morel, "A Review of Image Denoising Algorithms, with a New One," *Multiscale Modeling & Simulation*, vol. 4, no. 2, pp. 490–530, 2005. DOI: 10.1137/040616024.

[17] Z. Jian, L. Yu, B. Rao, B.J. Tromberg, and Z. Chen, "Three-Dimensional Speckle Suppression in Optical Coherence Tomography Based on the Curvelet Transform," *Optics Express*, vol. 18, no. 2, pp. 1024–1032, Jan. 2010. DOI: 10.1364/OE.18.001024. PMID: 20173923; PMCID: PMC2898712.

[18] S. Adabi, Z. Turani, E. Fatemizadeh, A. Clayton, and M. Nasiriavanaki, "Optical Coherence Tomography Technology and Quality Improvement Methods for Optical Coherence Tomography Images of Skin: A Short Review," *Biomedical Engineering and Computational Biology*, vol. 8, pp. 1179597217713475, Jun. 2017. DOI: 10.1177/1179597217713475. PMID: 28638245; PMCID: PMC5470862.

[19] Fang, Y., Shao, X., Liu, B., Lv, H. (2023). Optical coherence tomography image despeckling based on tensor singular value decomposition and fractional edge detection. Heliyon, 9(7), e17735. doi: 10.1016/j.heliyon.2023.e17735. PMID: 37449117; PMCID: PMC10336597.

[20] D. Cui, E. Bo, Y. Luo, X. Liu, X. Wang, S. Chen, X. Yu, S. Chen, P. Shum, and L. Liu, "Multifiber angular compounding optical coherence tomography for speckle reduction," *Opt. Lett.*, vol. 42, no. 1, pp. 125–128, Jan. 2017. DOI: 10.1364/OL.42.000125.

[21] A. Ozcan, A. Bilenca, A.E. Desjardins, B.E. Bouma, and G.J. Tearney, *Speckle reduction in optical coherence tomography images using digital filtering*, J Opt Soc Am A Opt Image Sci Vis, vol. 24, no. 7, pp. 1901–1910, Jul. 2007, doi: 10.1364/josaa.24.001901, PMID: 17728812, PMCID: PMC2713058.

[22] Wang, Y., Liu, S., Lou, S., Zhang, W., Cai, H., & Chen, X. (2019). Application of optical coherence tomography in clinical diagnosis. *Journal of X-Ray Science and Technology*, 27(6), 995-1006. doi:10.3233/XST-190559

[23] M. Gargesha, M.W. Jenkins, A.M. Rollins, and D.L. Wilson, *Denoising and 4D visualization of OCT images*, Opt Express, vol. 16, no. 16, pp. 12313–12333, Aug. 4, 2008, doi: 10.1364/oe.16.012313, PMID: 18679509, PMCID: PMC2748663.

[24] J. Doustar, T. Torbati, K.L. Black, Y. Koronyo, M. Koronyo-Hamaoui, *Optical Coherence Tomography in Alzheimer's Disease and Other Neurodegenerative Diseases*, Frontiers in Neurology, vol. 8, article 701, Dec. 19, 2017. doi: 10.3389/fneur.2017.00701. PMID: 29312125; PMCID: PMC5742098.

[25] A.L.M. Almeida, L.A. Pires, E.A. Figueiredo, L.V.F. Costa-Cunha, L.C. Zacharias, R.C. Preti, M.L.R. Monteiro, L.P. Cunha, *Correlation between cognitive impairment and retinal neural loss assessed by swept-source optical coherence tomography in patients with mild cognitive impairment*, Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, vol. 11, pp. 659–669, Sep. 25, 2019. doi: 10.1016/j.dadm.2019.08.006. PMID: 31667327; PMCID: PMC6811896.

[26] A. Gheorghe, L. Mahdi, O. Musat, *AGE-RELATED MACULAR DEGENERATION*, Romanian Journal of Ophthalmology, vol. 59, no. 2, pp. 74–77, 2015. PMID: 26978865; PMCID: PMC5712933.

[27] S.J. Bakri, M. Bektas, D. Sharp, R. Luo, S.P. Sarda, S. Khan, *Geographic atrophy: Mechanism of disease, pathophysiology, and role of the complement system*, Journal of Managed Care & Specialty Pharmacy, vol. 29, no. 5-a Suppl, pp. S2–S11, May 2023. doi: 10.18553/jmcp.2023.29.5-a.s2. PMID: 37125931; PMCID: PMC10408405.

[28] Shin S, Sharma U, Tu H, Jung W, Boppart SA. *Characterization and Analysis of Relative Intensity Noise in Broadband Optical Sources for Optical Coherence Tomography*. IEEE Photonics Technology Letters. 2010;22(14):1057-1059. doi: 10.1109/LPT.2010.2050058. PMID: 22090794; PMCID: PMC3214975.

[29] Chen Y, de Bruin DM, Kerbage C, de Boer JF. *Spectrally balanced detection for optical frequency domain imaging*. Optics Express. 2007 Dec 10;15(25):16390-16399. doi: 10.1364/OE.15.016390. PMID: 19550929.

[30] N. Serrano and A. Bellogín, *Siamese neural networks in recommendation*, Neural Comput & Applic, vol. 35, pp. 13941–13953, 2023. doi:https://doi.org/10.1007/s00521-023-08610-0

[31] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," *Mechanical Systems and Signal Processing*, vol. 151, p. 107398, 2021.