MAIA

ERASMUS MUNDUS

JOINT MASTER
IN MEDICAL IMAGING
AND APPLICATIONS

Joint Master in Medical Imaging and Applications

Master Thesis Proceedings

Promotion 2016-18

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; 2018).

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.

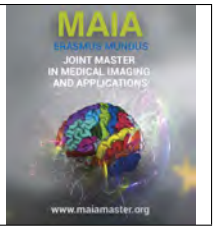
We finally want to thank and congratulate all the students for their effort done during this last semester of the Joint Master Degree in Medical Imaging and Applications.

MAIA Master Academic and Administrative Board

Contents

Medical Imaging and Applications

Master Thesis, June 2018



Malignancy estimation of Pulmonary Nodules using Multi-View Multi-Time Point Convolutional Neural Networks

Tajwar Abrar Aleef, **Supervisors**: Colin Jacobs & Bram van Ginneken

Radboud University Medical Center, Nijmegen, Netherlands

Abstract

Lung Cancer is one of the leading causes of cancer-related deaths for both men and women in the United States. The aim of lung cancer screening is to detect lung cancer at an early stage. Majority of the time, after the lung nodule detection phase, only a small portion out of all the nodules that get detected turns out to be cancerous. Compared to traditional techniques that use handcrafted features and furthermore relies on tedious & time-consuming prior lung nodule segmentation, the proposed method uses deep learning techniques in an end-to-end arrangement that performs both the feature extraction and classification directly from raw nodule patches. In this study, we focus on improving the pulmonary nodule malignancy estimation part by introducing a novel multi-view multi-timepoint convolutional neural network (MVMT-CNN) architecture that uses low dose CT images as its input. The dataset used in this study was taken from the National Lung Cancer Screening Trial (NLST)- which is the largest lung cancer screening trial known to date. We investigate the influence of whether adding temporal information of the same patient can help to improve the diagnosis. The proposed convolutional neural network architecture requires nine 2D patches- each of which represents a certain plane from the extracted 3D nodule patches. The nine planes are analyzed separately in parallel CNN streams and the output features coming from the nine different pathways are fused into one layer before passing it to the classification stage. Additionally, batch normalization and drop out layers are also incorporated in order to decrease the training time and reduce the chances of over-fitting. The average Area Under the ROC curve obtained after 5 fold cross validation along with bootstrapping were used to compare & select the final best performing architecture. The robustness of the final selected model was examined and verified by swapping the time points to see if the network did actually learn to identify the growth of the nodule between timepoints. The proposed method confirms that using the proposed multi-view multi-timepoint CNN architecture improves the prediction ability of pulmonary nodules significantly.

Keywords: Pulmonary nodule, lung cancer, nodule malignancy, temporal data, Convolutional Neural Network (CNN)

1. Introduction

Lung cancer is the second-most common type of cancer diagnosed in the United States and it leads to the the most cancer-related deaths for both men and women. In 2018 alone, approximately 234,030 new cases of lung cancer are set to be diagnosed in the US. Additionally, it is expected that 154,050 deaths will occur due to lung cancer which will account for 1 out of every 4 cancer-related deaths in the US in 2018 (Society, 2018). Early detection as well as accurate localization of the nodules however can aid in increasing the survival rate of lung cancer up to 52% (Liu et al., 2018). Lung nodules which are commonly spherical in shape can be difficult to detect due to having surrounding anatomical structures such vessels and pulmonary walls (Hussein et al., 2017). From the detected nodules, only around 20% turns out to be cancerous (Erasmus et al., 2000). Considering the immense variations of lung nodules, even experienced radiologists can fail to correctly identify the cancerous nodules. For that reason developing robust automatic distinction systems for lung nodules is a critical step

for both screening and clinical use in the diagnosis of lung cancer. Back in the early 2000's, research showed that it is possible to decrease the mortality rate of lung cancer with low dose CT scans as CT scan produces 3D images with more finer details compared to standard Chest X-rays. This helps in improving the detection capability of lung nodules at early stages and hence allows for better treatment options (Henschke et al., 1999). The voxel values in CT scans represents the radiodensity of the tissues in Hounsfield scale (HU). The radiodensity of air and distilled water at standard pressure temperature is defined at -1000 HU and 0 HU. Lung is therefore perfectly suited for CT imaging as it mostly consists of air of radiodensity around -100 0HU and surrounding tissue density of around 0 HU. This considerable difference allows acquisition of high quality images with high contrast even with at a low radiation setting. This is very convenient, especially for cases with large screening programs where the participants are expected to be exposed to frequent CT scannings. Technological advances in CT image acquisition has made it possible to capture lung nodules and its surrounding tissues with intricate details. This particular increase in the quality of CT images has created a pathway for data-driven analysis for accurately detecting lung nodules and predicting their probability of malignancy, which in return results in better management decisions and effective development of lung cancer screening programs. The principal goal of lung cancer screening programs is to detect lung cancer at an earlier stage, during which the treatment and prognosis options are better. The stage at which the lung cancer is diagnosed dictates the treatment options available and it directly correlates to the mortality rate of the patient. 57% of all lung cancer diagnosis is done during the later stages as the symptoms such as repeated coughing, pain in the chest, blood in the sputum and reoccurring pneumonia are typically observed only at a later stage when the metastasis has already begun and cancer has grown by several centimeters (Ellis and Vandermeer, 2011). The 5 year survival rate for late detection is 4.5%, whereas, an early detection can raise the 5 year survival rate up to 55%.

The National Lung Screening Trial (NLST) study was conducted in the United States which included 26,722 patients who were scanned using low-dose CT and 26,744 patients who were scanned using Chest X-rays. This is the largest randomized screening trial conducted to investigate the benefits of using low dose CT over using conventional Chest X-rays for early detection of lung cancer. The participants who took part in this study were required to have at least a smoking history of a minimum of 30 packs per year. They could be both former or current smokers but they cannot have symptoms or any family history of lung cancer. Three screenings were conducted on all participant with an interval period of one year between scans and later they were followed up after five years. After randomization, nodules that were malignant were confirmed by biopsy up to 7 years after the initial randomization. From the study, the conclusion was that using low dose CT compared to traditional radiography decreased mortality due to lung cancer by 20% (Team, 2011b).

One big problem with such large lung screening trials is that it generates large quantities of data that must be analyzed one by one by the radiologists which is both expensive and time consuming. One of the main aims of this thesis is to develop a malignancy estimation system for lung nodules using a machine learning approach with the goal to significantly automate and reduce the workload of the radiologists. Current convention during the screening of lung cancer include the use of Lung Imaging Reporting and Data System (Lung-RADS). Lung-RADS is a management and nodule scoring system developed by the American College of Radiology (ACR) in order to standardize the follow-up steps of screening protocols (of Radiology, 2014). Lung-RADS has a set of categories, and each category is dependent on the nodule type, nodule size, and the growth rate. Each category then determines a follow-up recommendation which can be to take a new CT scan after a certain time period or to seek additional imaging techniques or to directly go for a biopsy. Each category also gives an estimate of the probability of the nodules being malignant.

Automatic prediction of nodule malignancy generally follows a common pipeline. First, the suspicious candidates that can be nodules are selected from the lung CT scans. Here, morphological operations are used to detect a huge number of candidates usually considering a high sensitivity. Next, a false positive reduction step separates the nodules from the non-nodules. After that, some system includes a segmentation step that separates the nodules from the background in order remove unnecessary information. Handcrafted features are then extracted from the region of interest followed by a classifier, which is trained to estimate the final malignancy of the nodules. Typical handcrafted feature extraction methods make use of histograms (Uchiyama et al. (2003)), scale invariant feature transform (SIFT) (Farag et al. (2011)), local binary patterns (LBP) (Sorensen et al. (2010)) and histogram of oriented gradients (HOG) (Song et al., 2013). The extracted features then are fed into classifiers such as Support Vector Machines(SVM) (2015) and Random Forests (Ma et al., 2016). Other methods include Zinovev et al. (2011), who used a belief decision tree method to differentiate the nodules semantic features. Chen et al. (2011) suggested an approach using an ensemble scheme of neural networks to predict the classes of nodules to benign, malignant and uncertain. Han

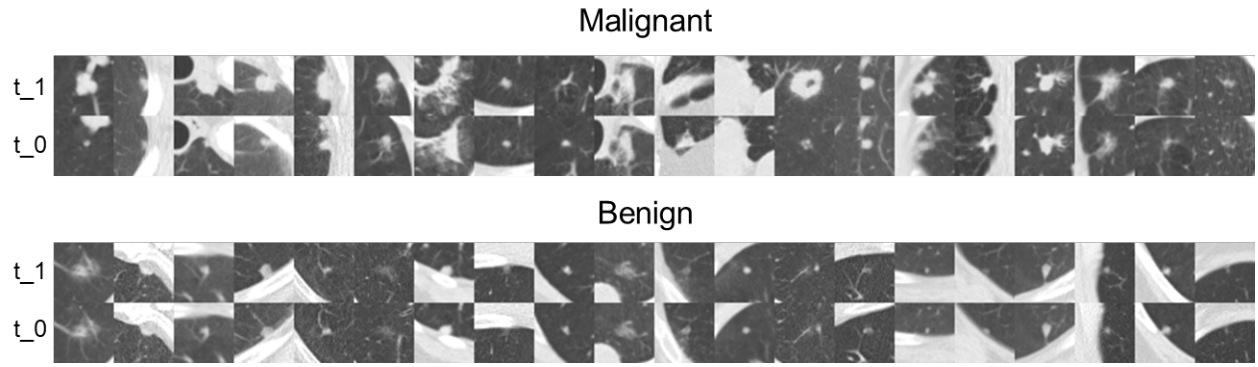


Figure 1: Figure showing two consequent timepoints (t_1,t_0) of malignant and benign nodule patches taken over 1 year of interval. From the patches, is can be seen that between the timepoints, the benign nodules does not change a lot. For the case of malignant nodules, a significant change/growth is seen for majority of the cases

et al. (2015) proposed a 3D image based texture feature analysis system for the purpose of nodule classification. Balagurunathan et al. (2014) and Aerts et al. (2014) worked on extracting features from nodules to study how efficiently they can be used for lung nodule malignancy prediction. Nowadays, with the availability of huge datasets along with the capability of massive parallelization permitted by modern GPU's, Convolutional Neural Networks are outperforming all existing state of the arts in both field of computer vision applications and medical imaging. Simonyan and Zisserman (2014),Brosch et al. (2014) and Parveen and Kavitha (2012) did a review on CAD systems designed for lung nodule analysis which includes preprocessing, segmentation and classification. The paper concluded that the use of artificial neural networks increases efficiency in lung nodule analysis. Hua et al. (2015) analyzed the use of deep learning techniques including Deep Belief Network (DBN) and Convolutional Neural Networks for diagnosis of lung nodules where they show that deep learning architectures is capable of beating any conventional method that uses handcrafted features. Lo et al. (1995) worked with pulmonary nodule detection in chest CT images using Convolutional neural network architecture.

Traditional studies commonly depend on careful prior nodule segmentation and tedious feature extraction before the classifier can be trained. Classical segmentation methods such as region growing which depends largely on the initialization threshold can lead to inaccurate segmentation and hence inaccurate feature extraction. In this study, an architecture is designed based on using the whole nodule patches directly as the input and then using an end-to-end convolutional neural network architecture to predict the malignancy of the nodules directly from the input patches of the nodules. In this thesis, the use of temporal information is studied intensively to verify if adding additional scans of the same nodules spanned over time can improve the performance of the output classification.

Prior to this work, to our best knowledge, no system exists that uses temporal data to predict malignancy of lung nodules. Our proposed model was obtained after extensive experiments and comparison between different models was made using data taken directly from the NLST dataset for testing and validation. A separate set also from the NLST dataset was kept on hold for testing the final model. Further tests are performed to conclude the robustness of the proposed system in nodule malignancy prediction. After careful comparison between different techniques, the proposed algorithms shows that using multi-time point scans in lung nodule malignancy prediction outperforms the current state of the art results of using only single time point data.

2. State of the art

As the main goal of having lung cancer screening programs is to identify the cancerous nodules at an early stage, an accurate nodule malignancy estimation system becomes a key part of having an efficient screening program. In spite of that, very few research has been conducted on developing lung cancer malignancy prediction systems using machine learning approaches. Commonly, a lot of studies can be found including open challenges that tackle the problem of nodule type detection-which is also a key part of the lung cancer diagnosis systems. The reason that there is a scarcity of research in nodule malignancy detection can be explained by the lack public data with pathologically proven malignant nodules. Public challenges like Luna16 and 2017 Kaggle Bowl challenge offered public dataset of lung nodule type classification with annotation. Some of the previous works included the use of linear discriminant analysis of the features extracted using morphology and gray-levels from lung nodules that were segmented using multi-level thresholding (Armato et al., 2003). Zinovev et al. (2011) distinguish between cancerous and no cancerous nodules by using both texture and intensity features with the help of belief decision trees and

a multi-labeling approach. Way et al. (2009) again did segmentation of the lung nodules by using k means clustering algorithm. Next, they used linear discriminant analysis again with surface, texture and morphological feature. It should be noted that most of the features used in such systems such as shape and volume are highly dependant on the segmented mast. This sensitivity arising from the segmentation phrase can hinder the classifier to learn the correct features for distinguishing cancerous nodules. Another big issue with such systems can be how to select the optimal set of features that best represent the discriminative characteristics of the lung nodules (Ciompi et al., 2015). Recently, with the use of deep learning techniques, such problems can be solved. Such systems are capable of automatically learning the best set of features from raw input images without ever needing prior sensitive segmentations. A few recent works have used deep learning techniques to classify nodule malignancy. Kumar et al. (2015) initially trained an unsupervised deep autoencoder that is able to extract complex unobservable features and then used decision trees to predict the malignancy of lung nodules. Hua et al. (2015) used a supervised approach with a deep belief network and convolutional neural network for estimating malignancy.Ciompi et al. (2015) used convolutional neural network models that were pre-trained for the purpose of classifying the nodules to peri-fissural nodules or nonperi-fissural. 2D patches of nodules in axial, sagittal and frontal planes were extracted from the CT image and this was fed into the pre-trained network. Next an ensemble of deep features along with a bag of frequency features was used to finally train the peri-fissural nodule classification system. Shen et al. (2015) used a multi-scale 3D CNN approach, where the input was 3D patches instead of the aforementioned 2D patch strategies.Shen et al. (2017) used the same theory of multiscaling but they showed a more efficient way of replacing the max-pooling layers with what they call a multi crop layer which is able to extract multiscale features without the need of having multiple parallel networks. The study also proposes to use the same architecture of multiscale multi-crop layers to estimate the diameter of the nodules which can further assist in predicting the malignancy of nodules. However, the use of such aggressive double max-pooling layers after the first convolution layer can hinder the ability to learn the spatial features correctly. Hussein et al. (2017) median intensity projection to obtain 2D patches in the axial, sagittal and coronal planes. The planes were then stacked in channels of a single tensor followed by CNN network that extracts the features of the nodules and finally, a Gaussian process regression makes the final classification.Nibali et al. (2017) used an architecture that was inspired by the original study of the deep residual network He et al. (2016) the proposed architecture is an 18 layer deep CNN architecture with skip connections between the layers like the original He et al. (2016) that won the ImageNet detection challenge. This paper also proposed the use of curriculum learning, transfer learning and experimenting with the network depth.

Due to a lack of a standardized publicly available dataset with proper annotations, different studies considered using different datasets for training and validation. It is, therefore, a difficult task to compare the methods and identify which one works best. Majority of the screening program performs follow up scans of the same patients during their study. Normally, the ability to predict the lung nodules with temporal information is more reliable due to having information of the growth of the nodules which is a direct predictor for nodule malignancy. To our best knowledge, none of the studies took advantage of using this temporal information to build a more robust lung nodule prediction system. The method proposed in this thesis shows how just by adding two timepoint information can significantly improve the prediction ability of lung nodules.

3. Material and methods

3.1. Pulmonary Nodules:

Commonly, lung cancers start as pulmonary nodules at its early stages which can be defined as having rounded opacity, being well or poorly defined and measuring up to 3 cm in diameter (Hansell et al., 2008). Pulmonary nodules can be of three main types: solid, semi-solid and non-solid nodules. For predicting the probability of malignancy of a nodule, parameters such as the type of nodule, the size of the nodule, the upper lobe location, morphology and emphysema scores are used (McWilliams et al. (2013), of Radiology (2014)). The result can only be validated and confirmed by performing a biopsy on sample tissue cells extracted from the cancerous region. Since only a small portion of all visible nodules ends ups as being diagnosed as cancerous, having a system that can predict the malignancy of such nodules with good precision using CT images would help in significantly reducing the number of biopsies required and would enable in a quicker diagnosis to be made which is essential for large screening programs.

3.2. Machine learning in Medical Imaging

Compared to humans, in order to comprehend images, computers require a set of discriminative features. Such discriminative features can be referred to as a feature vector of numerical information that the computer algorithm uses to classify and differentiate the input images. For instance, for predicting the malignancy of nodules, the nodule size, intensity and shape information can be used as the discriminative features. The machine learning algorithms generally learn to map the feature vector to the output labels that are provided by the human experts. However, designing such handcrafted methods which extract all the optimal features is more commonly arduous and they result in moderate performance. Instead of handpicking the features,

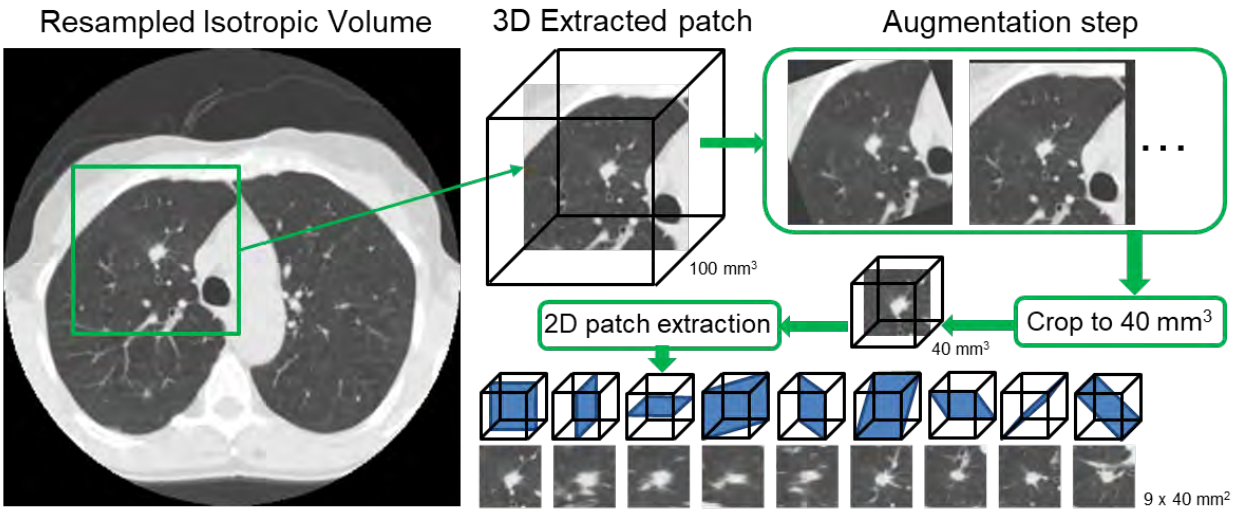


Figure 2: Pipeline from isotropic resampling, 3D patch extraction, augmentation, cropping and 2D 9 view patch extraction

another way of machine learning could be to allow the algorithm to extract the best set of features from raw images that most optimally distinguishes the dataset. In the machine learning universe, deep learning is one of such techniques that uses a set of feature extraction layers and classification layers to automatically extract features and perform classification directly from the input data (LeCun et al., 2015). Each consequential layer in a given deep learning model extracts features starting from simple features such as edges in the initial layers followed by more complex informations such as contours, shapes and more abstract features in the consecutive layers. Convolutional Neural Network is a subcategory of deep learning that uses the same idea of extracting features but now by using convolutional layers with receptive fields. The idea of convolutional neural networks was introduced back in the 1980s (Fukushima and Miyake, 1982). However, back in the days, due to the computational limitations and the lack of substantial amount of data, it was not popular and was not suited to be used in the medical field. With the improvements in available computing power and resources combined with the availability of large sets of datasets, now convolutional networks are widely becoming popular and are being used in all computer vision and medical image analysis tasks. They are consistently achieving state of the art performances and beating all the other gold standards that previously used handcrafted feature techniques. In this thesis, an end to end convolutional neural network system was proposed that can efficiently use the temporal information of lung nodules to generate reliable output predictions of the malignancy of lung nodules.

3.3. Dataset Preparation

The dataset used in this study was taken from the largest lung cancer screening trial known to date (Team, 2011a). The goal of the National Lung Screening Trial was to find out if low dose CT can improve the survival rates of lung cancer for a high risk group compared to a control group who was diagnosed using chest radiography. 26,722 participants in the NLST study was diagnosed using low dose CT across 33 US medical centers in the period between August 2002 to April 2004. These participants went through three screening rounds after randomization with a yearly interval and afterwards was followed up till December 2009. From the patients in the NLST study who received low dose CT, only a portion of the data was taken for this study which included participants who had high risk nodules, participants who passed away during follow up and a control group consisting of randomly selected subjects. The exact coordinates of the nodules present in the scans of the NLST dataset was not provided by the NSLT database. However, they do provide information such as the the number of nodules, the slice number and the lobar location. An in house software (CIRRUS Lung Screening, DIAG, RadboundUmc) was used to assist in detecting nodules that was missed in the NLST database. All nodules from the scans were evaluated and the malignant cases were identified by an experienced screening radiologist. For the cases in which the malignant nodules were visible and detected by the in house software in the prior scans. those nodules were also taken in the dataset. All the nodules for the patients who did not develop lung cancer were located by medical students who were trained by experienced radiologists. The nodules that could not be located were discarded from the dataset. The coordinates of the nodules that were annotated were not always centered and hence the annotation was just a rough approximate of the location of the nodules in the scan. From the total NLST database a subset of CT scans was selected that contained 218 malignant cases taken from 168 patients and

4229 benign cases taken from 2116 patients. From the selected dataset, some of the patients had nodules that were present in all three screening points and some had nodules that were visible only in two time points. In this study, only the last two time points were taken. The CT scans were in DICOM format which were first resampled to have an isotropic voxel spacing of 1mm x 1mm x 1mm. Next, for every nodule, using the annotated ground truth center coordinate information, a patch of $100mm^3$ cube was extracted taking 50 mm^3 on all 6 directions around the center coordinate. In cases were the nodules were too close to the boundary of the scan, voxels were padded with the minimum value of that corresponding patch. Figure 1 shows some examples of malignant and benign nodules extracted between the last two timepoints taken from the subset of NLST data. All the extracted 3D patches of nodules were saved to disk for further processing.

From the subset of the nodules extracted, 44 malignant nodules and 846 benign nodules were selected randomly and was kept in hold for being tested with the best performing model to check how well the system performs on unseen data. The rest of the 174 malignant and 3383 benign nodules were used to train and validate the system.

3.4. Data Augmentation

Normally, deep architectures of convolutional neural networks consists of a large quantity of trainable parameters. The larger the dataset, the better the network is able to learn the features and parameters which makes the model more robust and generalized in its output predictions. Trying to optimize a convolutional neural networks with a dataset that is not balanced between the classes can result in the network converging to a local minima where prediction is biased towards the class with the higher amount of data. Data augmentation allows a solution to this problem and also prevents the issue of over-fitting during training. The subset selected for training and validation had heavy imbalance between the two classes of nodules with 95% of them belonging to the benign class and the rest belonging to the malignant class. In order to balance the dataset, heavy augmentation was performed on the malignant nodules while preserving the semantic information of the nodules. Before augmentation, the dataset was divided into 5 sets of non-overlapping validation set made up of 20% of the total dataset. 5 sets were generated in order to allows a 5 fold cross validation during evaluating the performance of a given model. The validation sets were kept as they were and augmentation was performed only on the 5 training sets that was prepared by taking the rest of the data after excluding validation set for each of the 5 fold. During augmentation, Random rotations were applied on the extracted $100mm^3$ patches in the range between -20:20 degrees along with random translations in the range of -6:6 pixels for both x and y-axis. When performing random augmentation, both timepoints of every nodule received the same augmentation so that the correlation of the location of the nodules between the two scans are not hampered. Even though two scans coming from the two timepoints were not registered to each other, the annotations of the nodules allow extraction of the patches that are somewhat in the same space. The positive cases were augmented heavily until equal samples for both negative and positive cases were present for training. Since, if augmentation is only performed on a specific class, the network might learn the rotation/translation as a feature for distinguishing the two classes, similarly random rotation and translations were also applied on the original benign cases.

3.5. 2D patch extraction and Pre-processing

After augmentation, next step in the pipeline was to extract 2D multi-plane patches from the 3D patches extracted previously. From the center of the 3D patches, different planes of 2D patches were extracted, each with a voxel size of $40mm^2$. $40mm^2$ patch size was selected initially as statistically it was observed that 95% of the nodule diameters fell within 40mm, where enough contextual information about the nodules is available for the network to learn the correct features. Later during the study, a comparison is made between changing the size of the extracted 3D patches. Before extracting the 2D patches, the augmented original dataset was cropped from $100mm^3$ to $40mm^3$, after which the 9 planes are extracted for every nodule between the two-time points. As in Setio et al. (2016),Prasoon et al. (2013), (Roth et al., 2014) and (van Ginneken et al., 2015), the first three planes extracted are the commonly known axial, sagittal and coronal planes around the center of the cube. The other 6 planes are the diagonal planes that cut every corresponding parallel opposite face of the cube. Figure 2 shows the process of resampling, 3D patch extraction, patch augmentation, patch cropping and finally the 2D 9 view plane extraction.

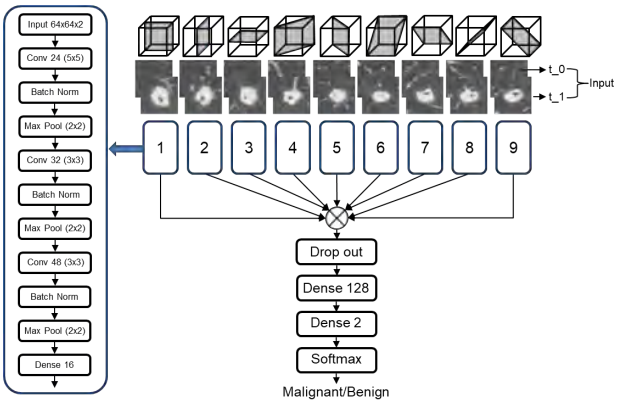


Figure 3: Baseline model with batch normalization and stacked time-series input

3.6. Convolutional Neural Network

The baseline network used in this study was inspired by the deep learning architecture used for false positive reduction in pulmonary nodule detection in Setio et al. (2016). The false positive reduction deep learning system was constructed out of a series of parallel streams of convolutional neural networks, where each of the stream takes a particular plane of the patch. This architecture is referred to as a multi-view CNN architecture as the network takes multiple 2D inputs, each coming from a different plane. The architecture in Setio et al. (2016) was developed by doing a study on a small dataset and then analyzing the most important hyperparameters that were critical to the output performance. Based on a pilot study using a smaller batch of data, they concluded that the number of planes and the type of fusion method used to merge the features taken from each parallel streams of the CNN architecture were the two most important hyperparameters. Considering the excellent performance demonstrated in the paper, it can be expected that the proposed architecture, that was designed for false positive reduction of lung nodules, was able to learn the latent discriminative features of pulmonary nodules. Keeping that in mind, a similar network is expected to work equally well in extracting features that can be used to design a classification system to find the malignancy of nodules. Hence, a baseline network was designed consisting of 9 parallel streams as introduced in the paper. Late fusion technique Setio et al. (2016) was used to fuse the features coming from the 9 different streams before feeding the concatenated features into the final classification layer of the architecture. For every nodule, the patches were resized from 40 x 40 pixels to 64 x 64 pixels using linear interpolation. Every 2D convolutional stream consisted of 3 convolutional layers. Starting with 24 filters of size 5x5 followed by 32 filters of size 3x3 and finally the last convolutional layer consisting of 48 filters of size 3x3. Even though batch normalization Ioffe and Szegedy (2015) is known to improve the performance, the original network did not have any batch normalization layers. Here in this study the network was evaluated with and without batch normalization. After every convolutional layer, a Batch normalization layer was added. Max pooling layer was added right afterwards the Batch normalizing layer with had a stride of 2. This reduces the size of the input patch to half of its input size by taking maximum values from a window size of 2x2. Rectified Linear units (ReLU) was used as the activation function Krizhevsky et al. (2012) for both the convolutional and fully connected layers. The activation function can be represented by the following formula:

$$a = max(0, x) \tag{1}$$

where, a is the output of the activation for a given input of x. The last fully connected layer gives an output of 16 neurons. The outputs from all the 9 streams are then concatenated using late fusion technique. Late fusion method (Prasoon et al. (2013),Karpathy et al. (2014)) is basically the method of fusing the outputs features of the parallel stream and outputting the features directly towards the final classification layers. The idea of adding the features of the 2D patches in such a way is to allow the network to learn the 3D characteristics from the individual set of features coming from every stream. For this setup, the weights between each parallel networks are shared which allows a reduction in the number of parameters that need to be learned from the dataset. A dropout layer Srivastava et al. (2014) was added which drops neuron connections with a probability of 50%. The output from the dropout layer was followed by two dense layers of 128 neurons and 2 neurons. Softmax activation was applied on the last layer to get the probability results for the binary classification problem.

3.7. Training Phase

Due to having a limited amount of data, the validation set contains only a small amount of data. Meaning that, based on the partition of the dataset, the validation results can have a significant variance. In order to somehow tackle that, 5 fold cross validation is used to evaluate the performance of the network. For every cross-validation, both the training and validation patches were normalized by using the following formula,

$$x = (x - mean)/std \tag{2}$$

where, mean and standard deviation are calculated from the training set used during a specific cross-validation. Stochastic Gradient descent was used to optimize the weights of the network with a starting learning rate of 10^{-3}. The learning rate was dropped by 20% after every 10 epochs. Cross-entropy error Nasr et al. (2002) between the predicted and ground truth was used as the loss function and the weights were updated using mini-batches of 64 training samples for every iteration. The initialization of the weights was done as proposed by Glorot and Bengio (2010) and the biases were initialized as zero. Training was continued for 50 epochs and only the weights for the highest AUC on the validation set were saved. The final evaluation was done by using all the 5 networks from the 5 fold training and averaging the results at the end.

The network was implemented using Keras with Tensorflow in the back-end.

3.8. Evaluation Metrics: ROC

The final motive of this thesis is to integrate the pulmonary nodule malignancy estimator with a fully fledged CAD system that takes a CT image as the input and outputs the location of the nodule with a malignancy score. Such systems have to be equivalent or superior to human observers in terms of both efficiency and accuracy of the results. The classifier proposed in this study gives an output value between 0 to 1, which basically

represents the probability of a nodule being malignant or not. An optimum threshold of the probability can be selected when computing the final binary result. A common way of visualizing the performance of such a binary classification system is to plot the Receiver Operating Characteristic (ROC) Curve, which is a plot of sensitivity on the y-axis (True Positive Rate) versus the 1-specificity on the x-axis (False Positive Rate) for every possible threshold on the classification probability. Sensitivity is the rate of positive samples that are correctly computed out of all the positive predictions. Similarly, specificity is the rate of true negatives that are correctly classified out of all the negative classifications. From the ROC curve, the area under the curve (AUC) can be calculated which determines the performance of the classification system. An AUC of 1 represents a perfect classification system where the True Positive rate is 1 and False Positive Rate is 0. The optimum threshold for a specific classification system can be selected by finding the point in the curve that is closest to the upper left of the ROC curve, where, the true positive rate is 1 and the false positive rate is 0.

3.9. Statistical Analysis

Machine learning algorithms make use of randomness. Especially in the case of deep learning, randomness can arise from the way the weights in the original network is initialized, randomness in the sampling/resampling of the data and randomness in the order the model sees the data. It is common to get different performance results on the same network even when trained using the same data. This uncertainty is of course bounded. However, knowing that such an uncertainty lies for every trained model, it is difficult to compare two methods relying only on a single evaluation. Although resampling methods of the dataset such as K fold cross validation help in reducing the uncertainty of the output performance and provide an average result of the performance that can be expected on the test set, it still can have a large variance. In order to evaluate the statistical significance between models bootstrapping method was used (Efron and Tibshirani, 1994). Bootstrapping allows a powerful and flexible statistical analysis that is used to quantify the ambiguity of a prediction model. 10,000 iterations were performed, where the performance of a model was evaluated by using the cross validation set that was used to train that particular model. For each iteration during bootstrapping, the corresponding validation set was randomly sampled with replacement following method from (Efron and Tibshirani, 1994). For every iteration, the performance metric was calculated and stored. Two main statistical analysis was done using the bootstrap method. The 95% confidence interval was estimated, which gives the range of performance values obtained during all the bootstrapping steps. Next, in order to compare two deep learning architectures, a statistical significance test was conducted. To compare two given systems, during each bootstrap sample the performance metric is evaluated for both systems and then is compared. This is done by calculating the p value, which is basically the number of times a systems performance was lower than the other system divided by the total number of bootstrap iterations.

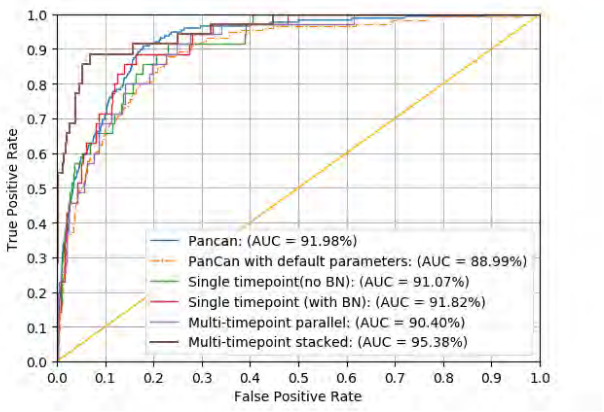


Figure 4: ROC of original PanCan, PanCan with default parameters, single timepoint with batch normalization, single timepoint without batch normalization, parallel multi-timepoint and stacked multi-time point

4. Results

4.1. Baseline Comparison model

4.2. PanCan model

Several well-known prediction models can be used to predict the probability of malignancy of nodules found in CT images. However, most of them are designed using a study based on a relatively small number of participants (Swensen et al. (1997), Gould et al. (2007), Herder et al. (2005)). In 2013, McWilliams et al. (2013) published a mathematical prediction model based on a large amount of screening data famously known as the Brock model or PanCan model. The model is called the PanCan model since it was developed using the PanCan screening data (McWilliams et al., 2013). The equation to estimate the probability of malignancy requires parameters such as the patient's age, gender, family history of lung cancer, emphysema score, nodule diameter, the location of the nodule, nodule count and spiculation. The model was validated using both internal and external screening data where it showed consistent performance on both the validation and external data. This is why this prediction model is recommended by many nodule management guidelines to be used as a primary tool for risk calculation for the case of both clinical use and screening trials (Callister et al. (2015),of Radiology (2014)). PanCan model was chosen as the baseline model for comparing the output of the proposed network during the analysis of each nodule. To obtain the parameters required to get the PanCan score, an experi-

enced radiologist was appointed to determine the nodule type, emphysema score and spiculation for each nodule. The family history, the age of the patients, gender of the patient, nodule count and lobe location of the nodules were obtained from the NLST pathology database.

The size of the nodules were estimated using a semi-automatic nodule segmentation tool Kuhnigk et al. (2006). As the PanCan model does not consider time series information, only the parameters from the last scan for every nodule was used to estimate the PanCan score. To evaluate the discriminative performance of the PanCan model, receiver Operation Characteristic (ROC) curve and the Area Under the ROC curve (AUC) was assessed as can be seen from Figure:4.

The following equation shows how to calculate the output probability of the PanCan risk model.

$$\begin{aligned}\textbf{Logodds} = {} & (0.0287 \times (Age - 62)) + Sex \\ & + FamilyHistory + Emphysema \\ & + Noduletype + NoduleLocation \\ & - (5.3854 \times ((Nodulesize/10)^{-0.5} \\ & - 1.58113883)) + Spiculation \\ & - (0.0824 \times (Nodulecount - 4)) \\ & - 6.7892\end{aligned} \tag{3}$$

$$\textbf{Cancerprobability} = 100 \times \frac{e^{\textbf{Logodds}}}{1 + e^{\textbf{Logodds}}} \tag{4}$$

Where, sex is equal to 0 for male or 0.6011 for female, family history is 0.2961 if there is family history of lung cancer, emphysema is 0.2953 if there is emphysema, nodule size is in mm, nodule type is -0.1276 if non-solid or 0.377 if partially solid or 0 is solid, nodule location is 0.6581 if located in upper lobe,

Similarly, a more simplified version of the PanCan model was used where parameters such age, sex, family history, emphysema score, nodule position, spiculation and nodule count was set to their default values for every nodule. For instance, the default value of age is 62, sex is male, family history is 0, emphysema is 0, nodule position is 0, nodule count is 1 and spiculation is 0. Hence, only the nodule size and nodule type were used for this simple logistic regression model. This was done as in most of the scenarios, all parameters are not available. Furthermore, features such as the nodule size and nodule type are considered more important when diagnosing the malignancy risk of nodules, so in order to compare the output scores, this simple regression model was also used. Figure 4 shows the ROC of the PanCan model with/without default parameters. It can be observed from the AUC that using all the parameters results in a better prediction. However, it should be noted that when using the default parameters, the AUC only falls by 3%. Meaning that the nodule size and nodule type information alone can differentiate the nodules with compelling results.

4.3. Single timepoint state of the art

The initial architecture design was based on Setio et al. (2016), where they used a multi-stream CNN architecture for false positive reduction during candidate selection of lung nodules. In the original work, they used input with no time series information. Hence, before feeding multiple time series information, the exact architecture was implemented. Here, only the last timepoint from every nodule was fed as the input for the model. As described before in section 3.7, 5 fold cross-validation was used to train and evaluate each model. This architecture was selected as the baseline CNN architecture, where the input 2D planes were extracted from $40mm^3$ patches containing the nodules. The results obtained from the baseline model is shown in Figure 4. The original architecture did not use batch normalization between the convolutional layers and max-pooling layers. Without Batch normalization, the average AUC for 5 fold cross validation was 91.09% with a standard deviation of 2.35%. The obtained AUC is comparable to the PanCan risk model with all parameters. The AUC is higher when compared to the PanCan model with default parameters. Next, Batch normalization was added after every convolutional layer and before activation function as in (Ioffe and Szegedy, 2015). The deep learning community has quickly adopted the use of batch normalization as is introduces some form of regularization which restrains the network from simply memorizing the training dataset, which means the network is expected to generalize on unseen data better with the use of batch normalization. Batch normalization also speeds up the convergence during training of neural networks. With batch normalization added, the AUC obtained was 93.46%(+/-1.30%). This experiment concluded that adding batch normalization improves the performance when dealing with nodule classification.

Using the exact same architecture with Batch Normalization, time series information was fed into the input. Two ways of giving the time series information were explored. The first method included two parallel branches inside the 9 streams, where each parallel branch processes input for a time series data and then at the end output a 16 neuron vector. The 16 neuron vector from the two time series for every parallel branch were fused into one and later the same classification layers were used as in the single timepoint architecture. Figure 6 shows the architecture of such parallel streams inside the 9 individual streams to separately process the time series input images. With such a strategy, the AUC obtained was 92.76%(+/-1.41%). The next strategy was to simply use the single timepoint architecture but instead of feeding a single image as input, here, the two-time points were stacked in channels of a single image and then fed as the input. Figure 3 shows the baseline model including batch normalization and stacked input images. With such a setup, the AUC obtained

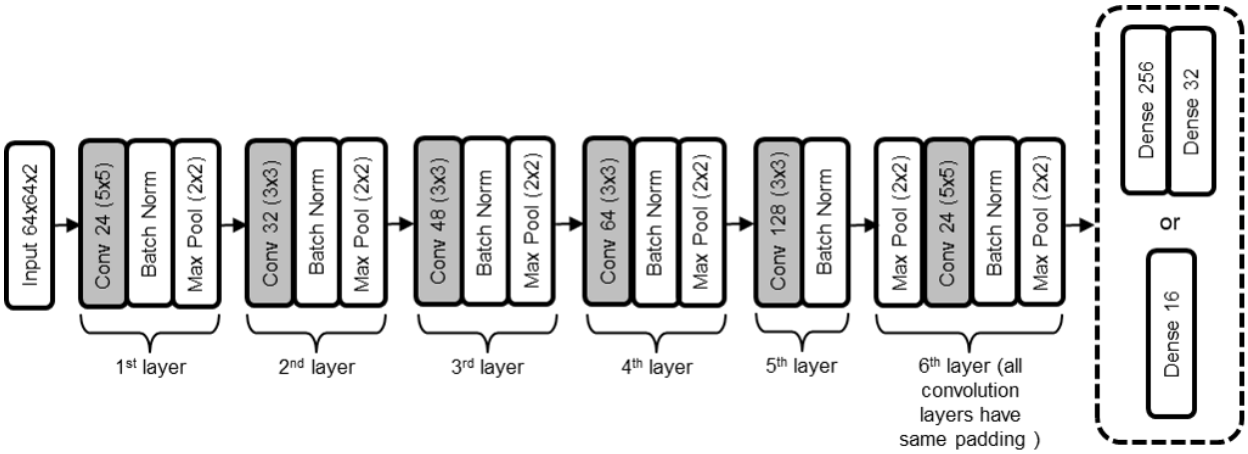


Figure 5: Schematic of the experiment by altering layer depth

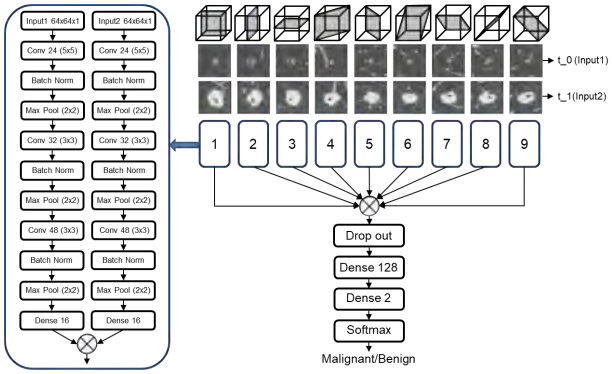


Figure 6: Parallel multi-timepoint architecture

was 95.69%(+/-1.48%). As seen from Table 1, the average AUC for the multi-time point network with parallel branch is less than that of using a single timepoint input. But stacking the timepoints as a single image increase the average AUC significantly. This is because not only the network needs less learnable parameters as can seen from Table 1, but also the network learns a better discriminative feature when stacked together in a single stream. Nodule diameter, which is an important feature in predicting the malignancy of a nodule, can easily be compared when time series information is available. Considering this, adding more time series data should improve the diagnosis decision. By processing the time series separately into two streams, the ability to compare the difference of the images from the input level is lost and only during the fusing of the feature vectors, the information of the two-time points are considered. If the parallel streams do not eventually learn that the growth of the nodule diameter is one of the substantial features, then the network does not fully utilize the time series information as can be seen from the results in Table 1. However, stacking the two images in channel gives an instant ability to compare the input raw images from the first convolution layer. This is why the results from stacking the images gives a better overall performance. From the series of experiments, it is clear that stacking time series information on the baseline model and adding batch normalization significantly improves the ability of the network to predict more accurately. The ROC figures for each of the different models are added in the Annex section of this report.

Table 1: Comparison of multi-timepoint and single-timepoint input data

Model	AUC(SD)	Parameter
PanCan with all parameters	91.98%	-
PanCan with default parameters	88.99%	-
Single Timepoint (no Batch Normalization)	91.09%(+/- 2.35%)	67,922
Single timepoint (with Batch Normalization)	93.46%(+/- 1.30%)	68,130
Two timepoints (in parallel channels)	92.76%(+/- 1.41%)	135,874
Two-timepoints (stacked in channels)	95.69%(+/- 1.48%)	68,730

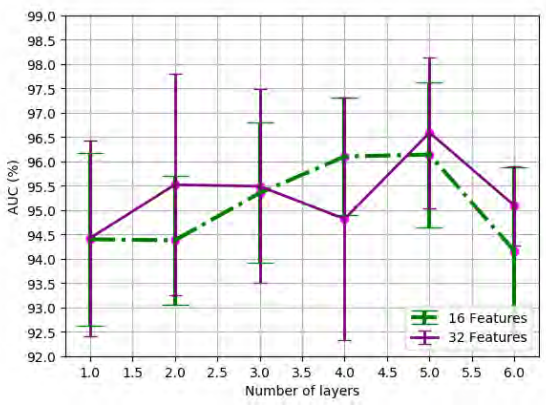
4.4. Selecting the layer depth

One of the leverage of deep learning networks is their ability to naturally fuse low, mid and high level features and the classifiers in a total end to end manner (Zeiler and Fergus,He et al. (2016)). Different features are learned in different levels of the stacked architecture and recent studies (Simonyan and Zisserman (2014),Srivastava et al. (2015)) suggests that deeper the architecture, the better the network is able to learn complex arbitrary features which in turns improves the output performance. This is also backed by the top scorers (Simonyan and Zisserman (2014),Szegedy et al. (2015)He et al. (2015),Ioffe and Szegedy (2015)) in the ImageNet challenge Russakovsky et al. (2015), where all of the architectures included deep stacks of layers. However, putting a lot of layers is not always the answer. Convergence can be affected by the vanishing gradient (Bengio et al. (1994),Glorot and Bengio (2010)), where, during backpropagation, the gradients get smaller and smaller as they progress backward from the last layer to the first. A deeper architecture could mean that by the time the gradients reach the earlier layers, they are so small that

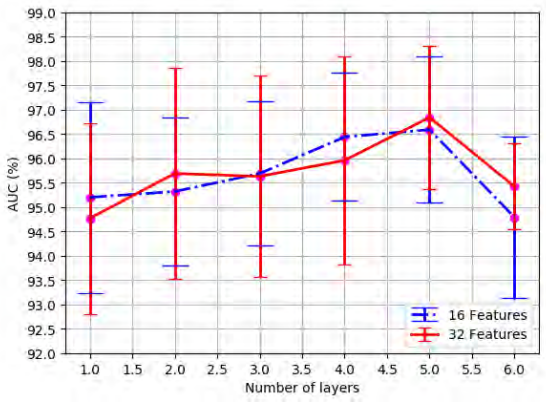
Table 2: Layer and output feature comparison

Layers	Average AUC: 16 features	95% CI: 16 features	p: 16 features	Parameters: 16 features	Average AUC: 32 features	95% CI: 32 features	p: 32 features	Parameters: 32 features
1	95.20%(+/- 1.96%)	92.6%-96.9%	0.0071	729,754	94.77% (+/- 1.96%)	91.5%-96.3%	0.0000	5,576,602
2	95.32%(+/- 1.51%)	92.4%-96.9%	0.0049	127,466	95.69%(+/- 2.17%)	93.3%-97.3%	0.0538	1,659,642
3	95.69%(+/- 1.48%)	93.9%-97.8%	0.1210	68,730	95.63%(+/- 2.07%)	93.1%-97.2%	0.0049	510,346
4	96.44%(+/- 1.31%)	94.3%-97.6%	0.1705	73,018	95.96%(+/- 2.13%)	92.6%-97.3%	0.0162	161,354
5	96.59%(+/- 1.50%)	93.7%-97.6%	0.0480	104,122	96.84%(+/- 1.47%)	94.9%-97.9%	-	161,738
6	94.79%(+/- 1.65%)	92.5%-96.7%	0.0013	238,010	95.42%(+/- 0.88%)	93.0% - 97.0%	0.0011	326,346

those layers learn very slowly compared to the last layers. So in order to find the optimum depth, the baseline model was tested with different combinations of network depth and output features. The classification layers ere kept same and only the parallel streams where altered to see which combination allows the network to learn the discriminative features from each plane of the nodule. The depth and output feature of the baseline model which had3 convolution layers and 16 neuron output was altered from having a single convolution layers to 6 convolutional layers. And for every change in layer, both 16 neuron and 32 neuron output features were experimented. For the 32 neuron output vector, an extra dense layer with 256 neurons was added preceding the 32 neuron dense layer. Figure 5 shows hows the architecture of each parallel stream looked like based on the number of layers and output feature selected. All the networks were trained using 5 cross-fold validation on the same NSLT dataset where the patch size of the nodules were $40mm^3$. 95% confidence interval was calculated using bootstrapping of 10,000 iterations on the validation set. From Table 2, it can be seen that the 5 layer 32 feature gives the best average AUC of 96.84%(+/- 1.47%) with the best confidence interval 94.9%-97.9%. The p value for all the other combinations were calculated by comparing all methods with the 5 layer 32 feature model. The p values from Table 2 shows that the selected 5 layer and 32 output neuron combination in fact performs significantly better than many other combinations. However, for model, the p value was still big enough to conclude if the 5 layer 32 model was significantly better or not. Hence, it only selected because it had the best combination of the average AUC, the 95% CI and the number of trainable parameters. Figure 7 shows how the output average AUC and its standard deviation changes with the layer depth. Figure 7(a) shows the change in average AUC when the model was optimized based on the lowest validation loss while Figure 7(b) shows the change in average AUC when the network is optimized based on the highest validation AUC. Figure 8 shows the ROC curves of models with different layer size for a specific validation set.



(a) AUC based on validation loss vs change in layers



(b) AUC based on highest validation AUC vs chnage in layers

Figure 7: Figures showing how the average AUC changes with the change in depth of a network.

4.5. What should be the optimum patch size:

Since the proposed method does not depend on any segmentation techniques prior to classifying the nodules, the field of view of the nodules that is presented to the deep learning network becomes an important factor. Since the size and shape of nodules vary quite significantly, a patch size that is too small can mean that for bigger nodules, important voxel values can go outside the patch. Again a larger patch size might introduce too much background information, where the network can learn features that are not locally from the nodules themselves. To test that, the optimum 5 layers 32 feature network was trained and validated using different size nodule patches. All the planes extracted from the patches were later resized to 64x64 pixels in order to fit the input size of the network. Table 3 shows the average AUC results from the 5 fold cross-validation. Similarly,

statical analysis was performed using bootstrapping of 10,000 iterations on 80% of the validation set. After bootstrapping the 95% confidence interval and p values of the different experiments were compared. The p values in Table 3, it is clear that there isnt a significant difference between using any of the patch sizes as p value was not below 0.05 for any of them. Hence, 40 mm^3 patch size was empirically chosen.

Table 3: Performance of the optimum model with different input patch size

Layer, Features, Patch size	**Average: AUC(+/- SD)**	**95% CI**	**p value**
5 layer, 32, 30mm^3	96.41%(+/-2.10%)	94.6%-97.8%	0.4483
5 layer, 32, 40mm^3	96.84%(+/-1.47%)	94.9%-97.9%	-
5 layer, 32, 50mm^3	95.86%(+/-1.80%)	93.3%-97.5%	0.2442
5 layer, 32, 60mm^3	96.05%(+/-1.12%)	94.0%-97.4%	0.2691
5 layer, 32, 70mm^3	95.83%(+/-1.75%)	94.0%-97.4%	0.2707
5 layer, 32, 80mm^3	94.55%(+/-2.18%)	91.8%-96.6%	0.0688

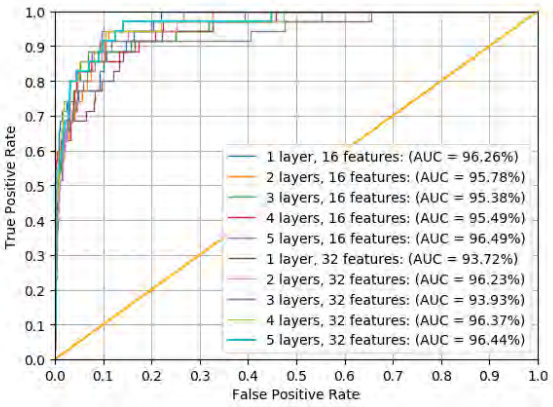


Figure 8: ROC of varying layer depth and output feature size

4.6. *How robust are the predictions?*

In order to check how robust the network is in predicting the proper discriminative features of the nodules, the trained network was put to the test first by checking what happens when the timepoints are swapped. As previously mentioned, the growth of nodule or change in diameter of the nodules is a crucial feature in understanding if a given nodule is malignant or not. For this experiment, the time points were swapped to see if the network is affected by it. With the swapped timepoints now the previously trained 5 layer 32 feature model was used to predict the AUC of the validation set. From the results a clear deterioration of the results is seen from the average AUC obtained from the 5 fold cross validation which fell from 96.84% to 76.92%. The AUC did not go down so much because swapping the benign nodules had little effect on the ability of the network to predict the benign class. This is because benign nodules have little to no change between the timepoints. However, when the timepoints of the malignant nodules were swapped, the true positive rate fell sharply as there is a negative growth in all the malignant nodules. This test confirms that the network learns that the increase in the size of the nodule is an important feature in distinguishing malignant nodules from benign ones. The accuracy of predicting positive cases only fell from 80% to 26%.

Since the time series scans taken in the NLST dataset are registered to some degree between the timepoints. For the next experiment, a question a raised that how well will the network perform if the two-time points are not registered like before. For this experiment, the augmentation was performed randomly between the timepoints, meaning that the same augmentation was not done on both the timepoints like previously. After 5 fold cross-validation, the average AUC was 96.47% (+/- 0.87%) (p=0.3004) with a 95% confidence interval of 94.3% and 97.4%. This shows that even the nodules that are not registered can be used for training as well as testing without significantly reducing the output performance of the system.

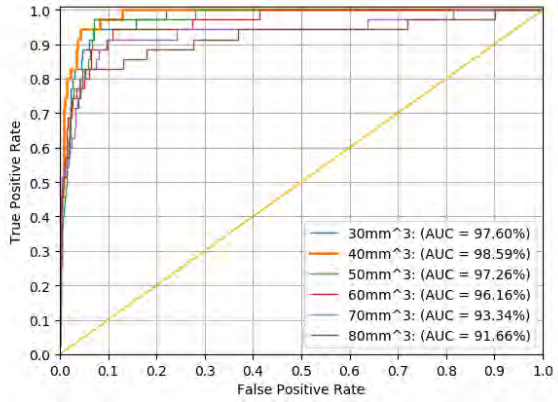


Figure 9: ROC of varying the 3D nodule patch size extracted from the original CT scan

4.7. *Evaluation on separate test set*

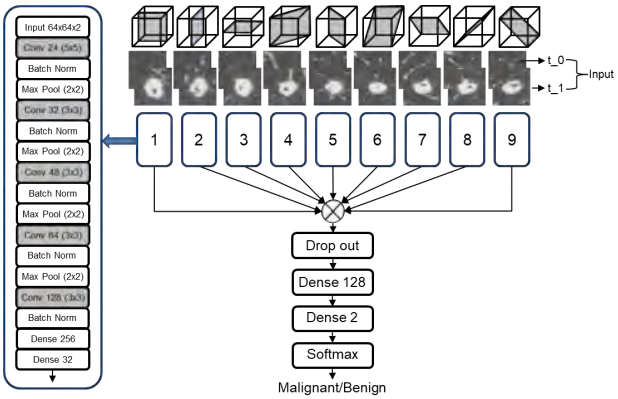


Figure 10: Architecture of the final 5 layer 32 feature model

The results of using the test set could not be obtained due to not having access to the ground truth annotations for the time being. The result is expected to be added in the final journal paper. For the scope of this thesis, only validation set was used to compare all the architectures. For the 5 layer 32 feature network (Figure 10), the operating point for each cross validation was calculated by finding the closest point of the ROC curves to the upper left hand corner of the ROC curve. The output prediction from the 5 models are then averages to get the final predictions. Figure 11 shows some of the misclassified benign and malignant nodules from the best performing model. By looking at the images and comparing between the timepoints, for the malignant cases it is seen that these particular nodules did not grow significantly. For the benign cases, a variation is seen between the timepoints which include background variation as well. This can be one of the reason why the nodules got misclassified.

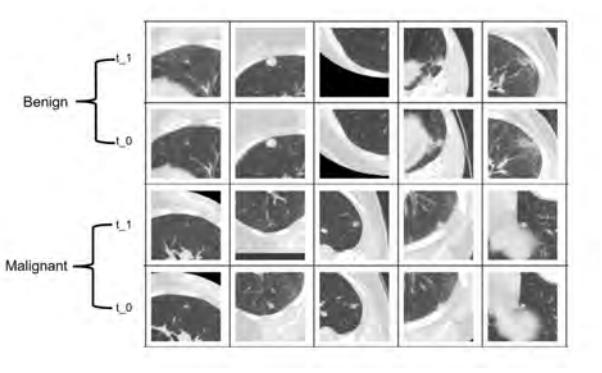


Figure 11: Misclassified results from the 5 layer 32 feature model

5. Discussion

In this study, a multi-view multi-timepoint convolutional neural network (MVMT-CNN) architecture was proposed for the purpose of estimating lung nodule malignancy using low dose CT images. Unlike conventional methods which require nodule segmentation and handcrafted features, the proposed system adopts an end-to-end classification method which learns the discriminative features of lung nodules automatically in an hierarchical manner. The goal of developing such a system is to aid in estimating the early suspiciousness of lung nodules using noninvasive CT images. This removes the need to surgically perform biopsy during early detection stages hence enabling a quick and cost-effective way of predicting suspiciousness of lung nodules. Is this thesis, the use of multi-time series information was studied extensively to justify if using temporal information of the same patients can aid in developing a better prediction system. The dataset used during training and validation of the various models were taken from the large lung cancer screening trial (NLST). For the first series of experiments, the main goal was to compare if adding additional time series information improves the results. Before diving into the deep learning architectures, a baseline prediction model that uses handcrafted parameters was investigated using the same dataset. As K fold cross-validation technique was used for evaluating the deep learning architectures, the models indeed were evaluated using all the dataset. This is because, during 5 fold cross-validation, all samples are divided into 5 validation sets. Similarly, using PanCan prediction model, the probability of malignancy were generated for all the nodules in the dataset using information of only the last timepoint since PanCan model was not developed to handle temporal information. A separate version of the PanCan model that uses only the default parameters was also evaluated. Next, a baseline model was designed based on Setio et al. (2016), where using both single time point and multi timepoint (with temporal data) were examined. This architecture takes different planes/views of the patch of the nodules as input. The idea behind using patches coming from nine planes is because using a singular plane can result in the network learning incorrect features that can come from the background information such as the pulmonary walls/vessels. Furthermore, adding the extra planes also captures the 3D characteristics of the nodules in separate 2D patches and allows the network to learn the morphological characteristics from a 2D point of view. This significantly reduces the complexity of the network and allows faster training and prediction times. This baseline architecture was examined with various ways of inputting the patches. From this series of experiments, it was clear that adding batch normalization and stacking time series data in channels of the input image improves the average AUC of the system from 91.09% to 95.69%. This is expected because as from the results from the PanCan prediction model, it is clear that even with default parameters, the performance does not go down by a lot. Meaning that only the nodule type and nodule diameter is used for computing PanCan risk model with default parameters, nodule size is an import factor to discriminate between the two classes. With temporal information, the nodule growth can be estimated as well as the nodule diameter. With this additional parameter, the network is expected to perform better. Next series of experiments hypothesized how changing the architecture affects the output results. Different combinations of depths of layers and output features of the parallel streams were assessed and statical significance test was conducted with bootstrapping to compare the models. As seen from resent Imagenet challenges, deeper architectures are performing better than all the current state of the art results in computer vision applications. By varying the depth and width of the network, the networks are most likely to learn deep features and make more correct assumptions (Krizhevsky et al., 2012). From this experiment it was seen that a combination of 5 layers of convo-

lution layer and 32 feature output resulted in the best average AUC and this was validated by finding the p value between this model and the other combinations. Next, a hypothesis was put to the test that if changing the original patch size could make any significant improvement in the results. The size of the 3D patches extracted of the nodules were altered from 30-80mm^3. Again with statistical analysis, it was seen that a size of 40mm^3 gives the best performance. A few other complex models were compared with the best performing model 7 and it was concluded that the best results that were obtained on this dataset can be achieved with the 5 layer 32 features model. From the p values in Table 4, it can be seen that some of the models do come close to the results obtained from the 5 layer 32 features model. The difference+multi scale model had the highest p value (p=0.4276) in the set of models experimented with. It is also seen that the number of trainable parameters required for the difference+multi scale model is almost 4 times more. It can be possible if a larger dataset is used, the model with the higher parameter can learn more strong discriminative features and result in a better performance. But based on the limited amount of data used in the study, the 5 layer 32 feature model gives the best performance with an average AUC of 96.84% (95% CI:94.9%-97.9%). Further data balancing and inclusion more data should improve the result even further.

6. Conclusions

In this study, we proposed a deep learning architecture called multi-view multi-timepoint convolutional neural network (MVMT-CNN) that uses temporal data to predict malignancy directly from raw CT patches of nodules. To our best knowledge, this is the first system where temporal data has been used in a lung nodule malignancy predictor system. This study compared different methods to find the best performing model on the given dataset. The performance observed in this study shows encouraging results of using temporal data with deep learning architecture for lung nodule risk estimation. Further extension of the thesis can be to experiment with 3D fully convolutional neural networks and to test the proposed system on a larger dataset with more time point data.

References

Aerts, H.J., Velazquez, E.R., Leijenaar, R.T., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., et al., 2014. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature communications 5, 4006.

Armato, S.G., Altman, M.B., Wilkie, J., Sone, S., Li, F., Roy, A.S., et al., 2003. Automated lung nodule classification following automated nodule detection on ct: A serial approach. Medical Physics 30, 1188–1197.

Balagurunathan, Y., Gu, Y., Wang, H., Kumar, V., Grove, O., Hawkins, S., Kim, J., Goldgof, D.B., Hall, L.O., Gatenby, R.A., et al., 2014. Reproducibility and prognosis of quantitative features extracted from ct images. Translational oncology 7, 72–87.

Bengio, Y., Simard, P., Frasconi, P., 1994. Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks 5, 157–166.

Brosch, T., Yoo, Y., Li, D.K., Traboulsee, A., Tam, R., 2014. Modeling the variability in brain morphology and lesion distribution in multiple sclerosis by deep learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 462–469.

Callister, M., Baldwin, D., Akram, A., Barnard, S., Cane, P., Draffan, J., Franks, K., Gleeson, F., Graham, R., Malhotra, P., et al., 2015. British thoracic society guidelines for the investigation and management of pulmonary nodules: accredited by nice. Thorax 70, ii1–ii54.

Chen, H., Wu, W., Xia, H., Du, J., Yang, M., Ma, B., 2011. Classification of pulmonary nodules using neural network ensemble, in: International Symposium on Neural Networks, Springer. pp. 460–466.

Ciompi, F., de Hoop, B., van Riel, S.J., Chung, K., Scholten, E.T., Oudkerk, M., de Jong, P.A., Prokop, M., van Ginneken, B., 2015. Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2d views and a convolutional neural network out-of-the-box. Medical image analysis 26, 195–202.

Efron, B., Tibshirani, R.J., 1994. An introduction to the bootstrap. CRC press.

Ellis, P.M., Vandermeer, R., 2011. Delays in the diagnosis of lung cancer. Journal of thoracic disease 3, 183.

Erasmus, J.J., Connolly, J.E., McAdams, H.P., Roggli, V.L., 2000. Solitary pulmonary nodules: Part i. morphologic evaluation for differentiation of benign and malignant lesions. Radiographics 20, 43–58.

Farag, A., Ali, A., Graham, J., Farag, A., Elshazly, S., Falk, R., 2011. Evaluation of geometric feature descriptors for detection and classification of lung nodules in low dose ct scans of the chest, in: Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on, IEEE. pp. 169–172.

Fukushima, K., Miyake, S., 1982. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition, in: Competition and cooperation in neural nets. Springer, pp. 267–285.

van Ginneken, B., Setio, A.A., Jacobs, C., Ciompi, F., 2015. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans, in: Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on, IEEE. pp. 286–289.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp. 249–256.

Gould, M.K., Ananth, L., Barnett, P.G., 2007. A clinical model to estimate the pretest probability of lung cancer in patients with solitary pulmonary nodules. Chest 131, 383–388.

Han, F., Wang, H., Zhang, G., Han, H., Song, B., Li, L., Moore, W., Lu, H., Zhao, H., Liang, Z., 2015. Texture feature analysis for computer-aided diagnosis on pulmonary nodules. Journal of digital imaging 28, 99–115.

Hansell, D.M., Bankier, A.A., MacMahon, H., McLoud, T.C., Muller, N.L., Remy, J., 2008. Fleischner society: glossary of terms for thoracic imaging. Radiology 246, 697–722.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE international conference on computer vision, pp. 1026–1034.

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

Henschke, C.I., McCauley, D.I., Yankelevitz, D.F., Naidich, D.P., McGuinness, G., Miettinen, O.S., Libby, D.M., Pasmantier, M.W., Koizumi, J., Altorki, N.K., et al., 1999. Early lung cancer action project: overall design and findings from baseline screening. The Lancet 354, 99–105.

Herder, G.J., van Tinteren, H., Golding, R.P., Kostense, P.J., Comans,

E.F., Smit, E.F., Hoekstra, O.S., 2005. Clinical prediction model to characterize pulmonary nodules: validation and added value of 18 f-fluorodeoxyglucose positron emission tomography. Chest 128, 2490–2496.

Hua, K.L., Hsu, C.H., Hidayati, S.C., Cheng, W.H., Chen, Y.J., 2015. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. OncoTargets and therapy 8.

Hussein, S., Gillies, R., Cao, K., Song, Q., Bagci, U., 2017. Tumornet: Lung nodule characterization using multi-view convolutional neural network with gaussian process, in: Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on, IEEE. pp. 1007–1010.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 .

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L., 2014. Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1725–1732.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

Kuhnigk, J.M., Dicken, V., Bornemann, L., Bakai, A., Wormanns, D., Krass, S., Peitgen, H.O., 2006. Morphological segmentation and partial volume analysis for volumetry of solid pulmonary lesions in thoracic ct scans. IEEE Transactions on Medical Imaging 25, 417–434.

Kumar, D., Wong, A., Clausi, D.A., 2015. Lung nodule classification using deep features in ct images, in: Computer and Robot Vision (CRV), 2015 12th Conference on, IEEE. pp. 133–138.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. nature 521, 436.

Liu, X., Hou, F., Qin, H., Hao, A., 2018. Multi-view multi-scale cnns for lung nodule type classification from ct images. Pattern Recognition .

Lo, S.C.B., Chan, H.P., Lin, J.S., Li, H., Freedman, M.T., Mun, S.K., 1995. Artificial convolution neural network for medical image pattern recognition. Neural networks 8, 1201–1214.

Ma, J., Wang, Q., Ren, Y., Hu, H., Zhao, J., 2016. Automatic lung nodule classification with radiomics approach, in: Medical Imaging 2016: PACS and Imaging Informatics: Next Generation and Innovations, International Society for Optics and Photonics. p. 978906.

McWilliams, A., Tammemagi, M.C., Mayo, J.R., Roberts, H., Liu, G., Soghrati, K., Yasufuku, K., Martel, S., Laberge, F., Gingras, M., et al., 2013. Probability of cancer in pulmonary nodules detected on first screening ct. New England Journal of Medicine 369, 910–919.

Moeskops, P., Viergever, M.A., Mendrik, A.M., de Vries, L.S., Benders, M.J., Išgum, I., 2016. Automatic segmentation of mr brain images with a convolutional neural network. IEEE transactions on medical imaging 35, 1252–1261.

Nasr, G.E., Badr, E., Joun, C., 2002. Cross entropy error function in neural networks: Forecasting gasoline demand., in: FLAIRS Conference, pp. 381–384.

Nibali, A., He, Z., Wollersheim, D., 2017. Pulmonary nodule classification with deep residual networks. International journal of computer assisted radiology and surgery 12, 1799–1808.

Orozco, H.M., Villegas, O.O.V., Sánchez, V.G.C., Domínguez, H.d.J.O., Alfaro, M.d.J.N., 2015. Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine. Biomedical engineering online 14, 9.

Parveen, S.S., Kavitha, C., 2012. A review on computer aided detection and diagnosis of lung cancer nodules. International Journal of Computers & Technology 3, 393–400.

Prasoon, A., Petersen, K., Igel, C., Lauze, F., Dam, E., Nielsen, M., 2013. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 246–253.

of Radiology, A.C., 2014. Lung ct screening reporting and data system lung-rads^{TM}. https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Lung-Rads .

Roth, H.R., Lu, L., Seff, A., Cherry, K.M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., Summers, R.M., 2014. A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 520–527.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115, 211–252.

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

Shen, W., Zhou, M., Yang, F., Yang, C., Tian, J., 2015. Multi-scale convolutional neural networks for lung nodule classification, in: International Conference on Information Processing in Medical Imaging, Springer. pp. 588–599.

Shen, W., Zhou, M., Yang, F., Yu, D., Dong, D., Yang, C., Zang, Y., Tian, J., 2017. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. Pattern Recognition 61, 663–673.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Society, A.C., 2018. Cancer facts and figures 2018. American Cancer Society .

Song, Y., Cai, W., Zhou, Y., Feng, D.D., 2013. Feature-based image patch approximation for lung tissue classification. IEEE transactions on medical imaging 32, 797–808.

Sorensen, L., Shaker, S.B., De Bruijne, M., 2010. Quantitative analysis of pulmonary emphysema using local binary patterns. IEEE transactions on medical imaging 29, 559–569.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15, 1929–1958.

Srivastava, R.K., Greff, K., Schmidhuber, J., 2015. Highway networks. arXiv preprint arXiv:1505.00387 .

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.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., et al., 2015. Going deeper with convolutions, Cvpr.

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

Team, N.L.S.T.R., 2011b. Reduced lung-cancer mortality with low-dose computed tomographic screening. New England Journal of Medicine 365, 395–409.

Uchiyama, Y., Katsuragawa, S., Abe, H., Shiraishi, J., Li, F., Li, Q., Zhang, C.T., Suzuki, K., et al., 2003. Quantitative computerized analysis of diffuse lung disease in high-resolution computed tomography. Medical Physics 30, 2440–2454.

Way, T.W., Sahiner, B., Chan, H.P., Hadjiiski, L., Cascade, P.N., Chughtai, A., Bogot, N., Kazerooni, E., 2009. Computer-aided diagnosis of pulmonary nodules on ct scans: Improvement of classification performance with nodule surface features. Medical physics 36, 3086–3098.

Zeiler, M., Fergus, R., . Visualization and understanding convolutional neural networks .

Zinovev, D., Feigenbaum, J., Furst, J., Raicu, D., 2011. Probabilistic lung nodule classification with belief decision trees, in: Engineering in medicine and biology society, EMBC, 2011 annual international conference of the IEEE, IEEE. pp. 4493–4498.

7. Annex

This annex includes extra experiments and ideas of more complex CNN architectures. The idea here was to experiment with different cases to see if the results of 5 layer 32 features MVMT-CNN architecture can further be improved. The next series of experiments were performed again using the same training and validation dataset prepared from the NSLT dataset. 5 fold cross validation was similarly used and the extracted nodule patches had a size of $40mm^3$. Statistical analysis was performed at the end to see if any such network would perform better than the best-selected model. From the series of experiments, it is clear that the number of trainable parameters will increase with the network complexity. Keeping that in mind as well as the nature of having limited dataset with pathologically proven malignant nodules, none of the following architecture could top the 5layer 32 feature network. Table 4 shows the difference in performance between the different models experimented in this part.

7.1. Parallel+Stacked Model

From the previous experiments between using temporal data and single data, it was clear that the network learns better discriminative features when accessible to temporal data. The single timepoint network also performed comparatively, meaning, single timepoint scans also contains enough contextual information that can be used to differentiate cancerous nodules from noncancerous ones. The assumption made here was that while the stacked input stream learns features comparing the two-time points, adding extra parallel streams of similar CNN architecture that analyzes the two inputs separately can learn from local features that can help in discriminating the nodules better. The outputs of the three streams are concatenated before passing on to the next stage. Figure 17 shows the schematics of adding two additional parallel CNN architectures that analyze the two time points separately as well.

The ROC curve is generated for every fold of cross validation and can be observed in Figure 12. The average AUC for this model was 95.74% (95% CI: 92.9%-97.1%, p = 0.0063).

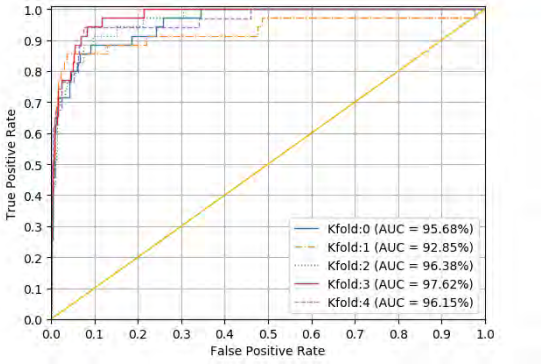
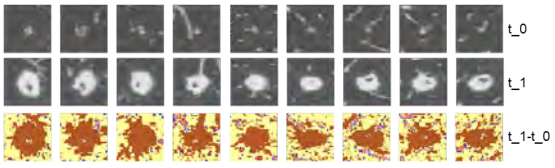


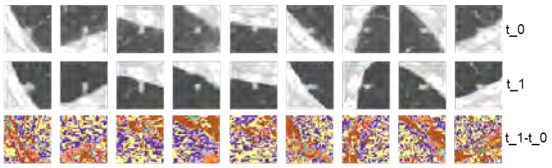
Figure 12: ROC of Stack + Parallel model

7.2. Difference Model

The main point of adding temporal data is to have a base of comparison between the change in the nodules characteristics over a period of time. Nodule growth which is an underlying predictor for nodule malignancy can be estimated by having time series information. Although it is expected that by stacking the images the initial convolution layers learns to map the difference of the inputs to the output label, here in this architecture, the difference between the two images is manually provided in a separate parallel branch. From figure 13, the benefits can clearly be seen. The figure shows the 9 views of a benign and a malignant nodule for two time points, t_0 and t_1. From the difference image it can clearly be seen that since the malignant cases are expected to grow, the difference around the center of each patch is high. Even if the patches are not aligned properly, the nodules are somewhat centered in the patch. Meaning, if there is a significant change in the shape of the nodule, a high difference should be observed in the middle portion of the patch. Similarly, since benign cases are not expected to grow, the difference image obtained by subtracting the benign cases gives less difference in the center portion of the patch.



(a) Malignant nodules



(b) Benign nodules

Figure 13: Figures showing how the difference images looks like for malignant and benign nodules in each of the 9 planes

Figure 18 shows the final architecture of the model with an additional branch in every stream which takes the difference image as the input. The ROC curve is drawn from the 5 cross-fold validation as can be seen in Figure 14. The average AUC for this model was 96.09% (95% CI: 93.9%-97.2%, p = 0.0786).

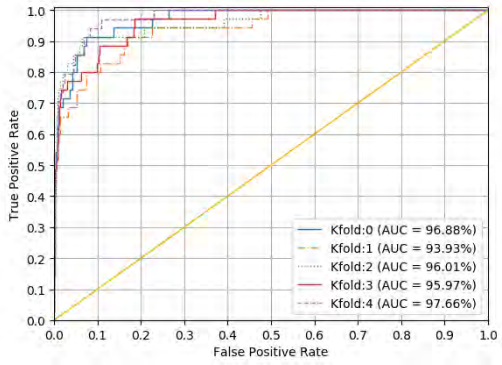


Figure 14: ROC of Difference model

7.3. *Multi-Scale Model*

For this experiment, the idea was that the size of nodule varies significantly in shape and sizes. Some nodules can be very small compared to the overall patch size. In order to extract the true local features coming from the nodules itself, the input needs to be cropped. Moeskops et al. (2016) shows that using a multiscale approach in classification task can allow the network to learn both local and global features which could benefit in the final classification task of the model. The stacked inputs where cropped from having 64x64 pixels to 32x32 pixels centered around the nodule and then it was passed through a parallel branch in every stream. With a smaller patch, the background information was significantly removed which meant now the network is now exposed to more voxels coming from the nodules itself. Since the patch size is reduced, the number of layers in the parallel branch processing the cropped images is reduced to only 2 convolution layers.

Figure 19 shows the architecture of the model using multi-scale information. The ROC curve is drawn from the 5 cross-fold validation as can be seen in figure 15. The average AUC for this model was 96.64% (95% CI: 94.1%-97.8%, p = 0.2611).

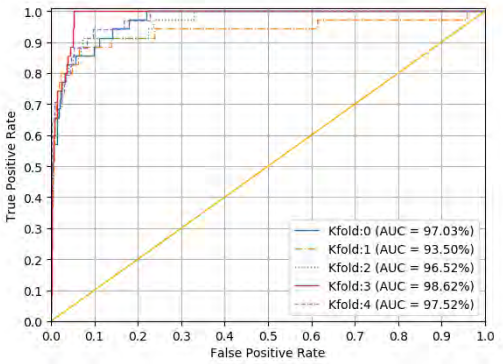


Figure 15: ROC of Multi-Scale

7.4. *Difference + Multi-Scale Model*

In this experiment, the idea was to combine both the difference branch and multi-scale branch with the best performing model. Figure 20 shows the architecture of the model using both the difference and multi-scale branch. The ROC curve is drawn from the 5 cross-fold validation as can be seen in figure 16. The average AUC for this model was 96.62% (95% CI: 94.8%-97.8%, p = 0.4276). The p value obtained is the highest p value obtained among any other experiments. Still for a bootstrap analysis of 10,000 iterations, only 4276 times, this model performed better than the previously determined best performing model. the conclusion meaning that the previously selected model still performs better than this model.

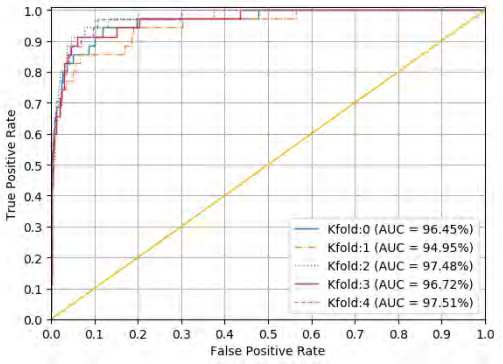


Figure 16: ROC of Difference + MultiScale

Table 4: Comparison of the experimented models

Model	Validation AUC	CI 95%	Parameters	p value
5 layer 32 feature	**96.84% (+/- 1.47%)**	**94.9%-97.9%**	**161,738**	**-**
Stacked + Parallel	95.74%(+/- 1.58%)	92.9%-97.1%	778,154	0.0063
Only Difference	96.09%(+/-1.25%)	93.9%-97.2%	519,098	0.0786
Only Multiscale	96.64%(+/-1.71%)	94.1%-97.8%	346,434	0.2611
Difference + Multiscale	96.62%(+/- 0.93%)	94.8%-97.8%	605,490	0.4276

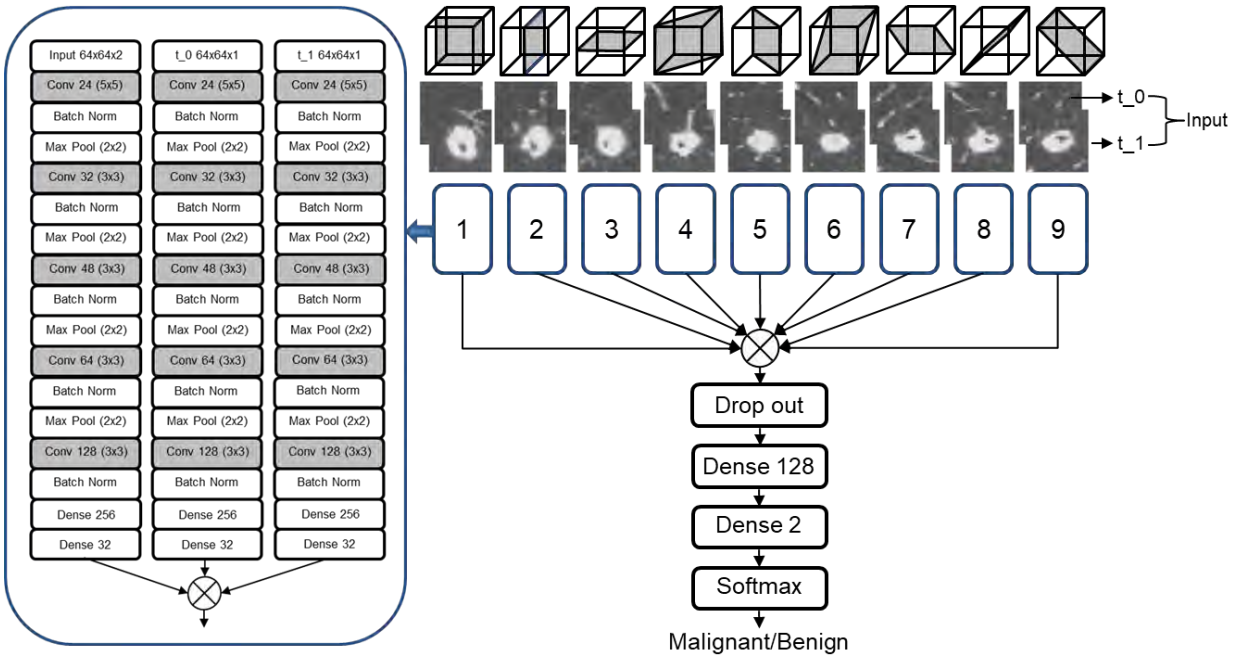


Figure 17: Architecture of Stacked + Parallel model

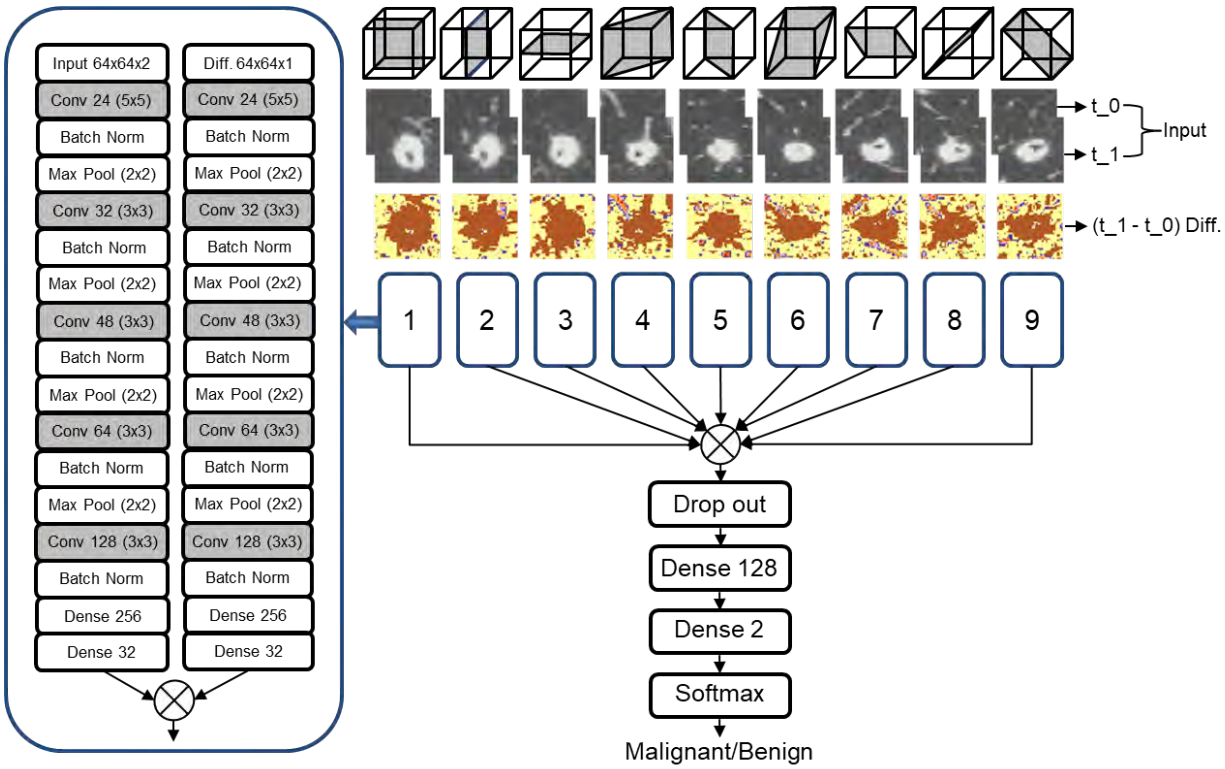


Figure 18: Difference model

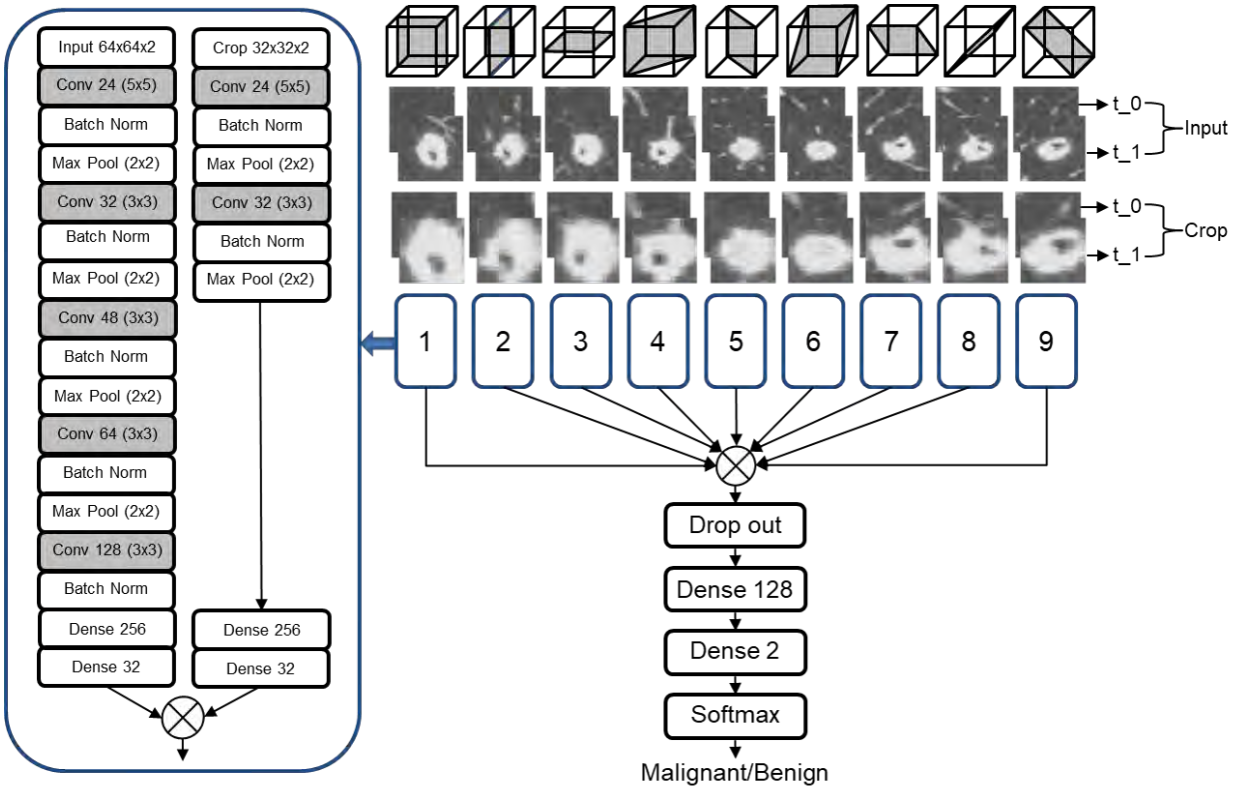


Figure 19: Multi-Scale model

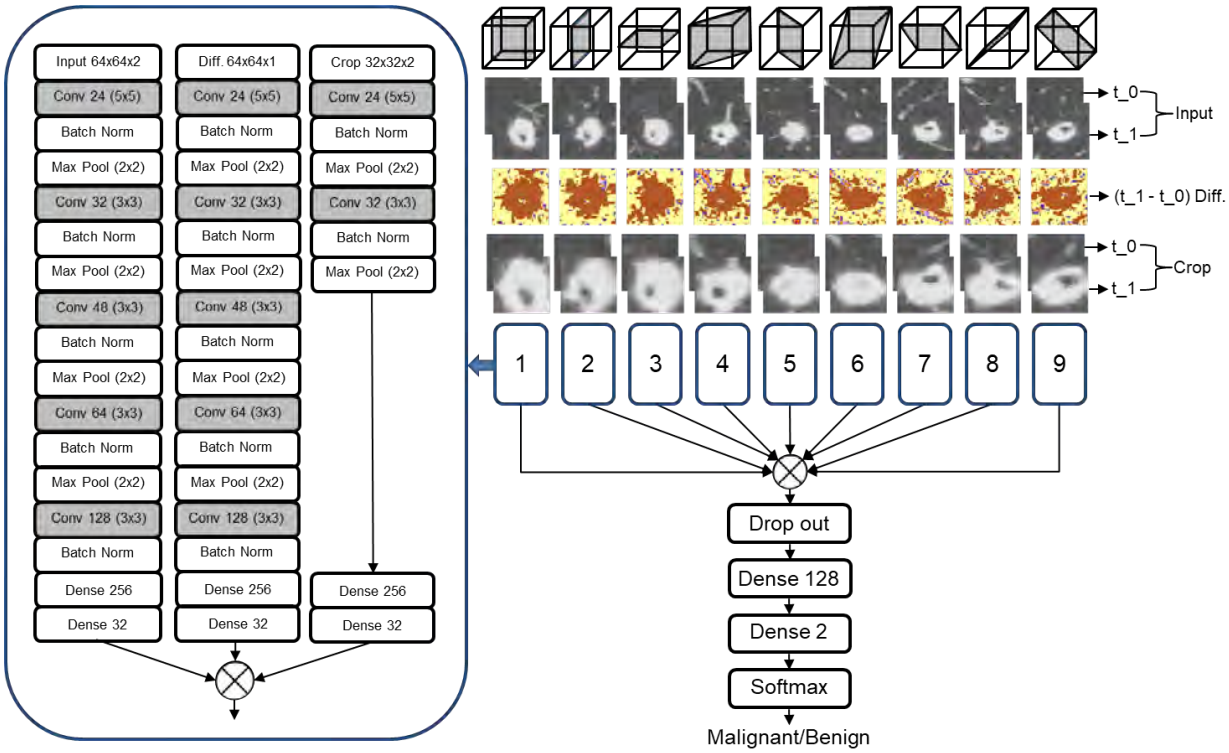
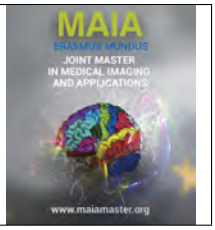


Figure 20: Difference + Multi-Scale Model

Medical Imaging and Applications

Master Thesis, June 2018



Characterisation and Matching of Skin Lesions through Deep Learning Techniques

Luca Canalini, Josep Quintana, Ricard Campos, Rafael Garcia

Computer Vision and Robotics Group, University of Girona, Catalonia, Spain
Coronis Computing, Girona, Catalonia, Spain

Abstract

The detection of new appearances and changes of pigmented skin lesions (PSLs) is essential for a timely diagnosis of possible developing melanoma. For this purpose, a common approach is to acquire baseline images of the skin surface and to use them as a reference to detect eventual changes in following examinations. To collect this baseline, total body skin examination (TBSE) is usually accomplished manually by a physician. However, the procedure can be time-consuming for patients with numerous skin lesions. In addition, it is highly prone to the subjectivity of the doctor performing the acquisition, increasing the risk of missing developing cancers. In this context, a second version of a new photogrammetry-based total body scanning system has been proposed to acquire skin images with better quality with respect to the first model. Additionally, given the state-of-the-art results obtained by deep learning techniques in image analysis, the existing pipeline has been replaced with a new one - based on artificial intelligence techniques - to automatically map and match skin lesions on the images acquired by the scanning system. The initial tests showed how the results improved the previous performance. The framework, indeed, can be applied in hairless areas of a patient to automatically detect and segment PSLs on full-body images. In addition, the paper proposes an automatic methodology for matching skin lesions within a scanning session.

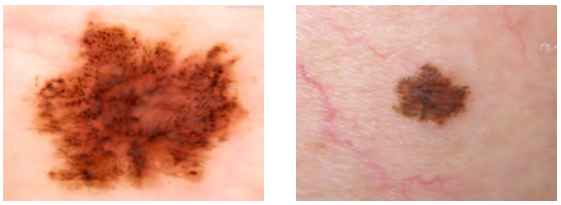
Keywords: Pigmented skin lesions, nevi, deep learning, total tody skin examination, scanner, melanoma

1. Introduction

Melanoma (also known as malignant melanoma) is a type of cancer that develops from the pigment-containing cells known as melanocytes and it is the most dangerous type of skin cancer[1]. Globally, in 2015 there were 3.1 million with people an active disease which resulted in 59,800 deaths (Stewart et al., 2017) and the country with the highest rates of melanoma is Australia and New Zealand. Nevertheless, timely diagnosis can prevent melanoma from producing any metastasis, so it can be cured completely (Weinstock, 2006).

Visual inspection is the most common technique for the diagnosis of melanoma (Wurm and Soyer, 2010), and the main characteristics that are taken into account for the diagnosis are the colour and the shape (Negin et al., 2003): lesions with irregularities in these two features are typically treated as candidates of possible developing cancer. Thus, physicians have to learn to recognise them. For this task, a popular method for remembering the signs and the symptoms of melanoma is the mnemonic acronym ABCDE (Friedman et al., 1985): **A**symmetrical skin lesions, **B**order of the lesion, **C**olor, **D**iameter, **E**nlarging. To understand whether a skin lesion would conduct to a possible melanoma, physicians need a baseline to compare it with respect to a previous state (Banky et al., 2005), to check if any relevant change has happened. One technique used to detect it is the total body photography (Halpern et al. (2003)), which consists in periodically acquiring images of the skin surface. Additionally, it is common practice to use the total body skin examination (TBSE) to compare skin lesions in a previous state to detect changes, achieving a more accurate diagnosis. However, acquiring this baseline is time-consuming and biased by the

[1] https://www.iarc.fr/en/publications/pdfs-online/wcr/2003/WorldCancerReport.pdf



(a) (b)

Figure 1: In situ melanoma acquired with two different modalities: in *a* a dermoscopy image, in *b* a clinical image of a lesion. The images were submitted to the Dermoscopy atlas by Dr. Alan Cameron

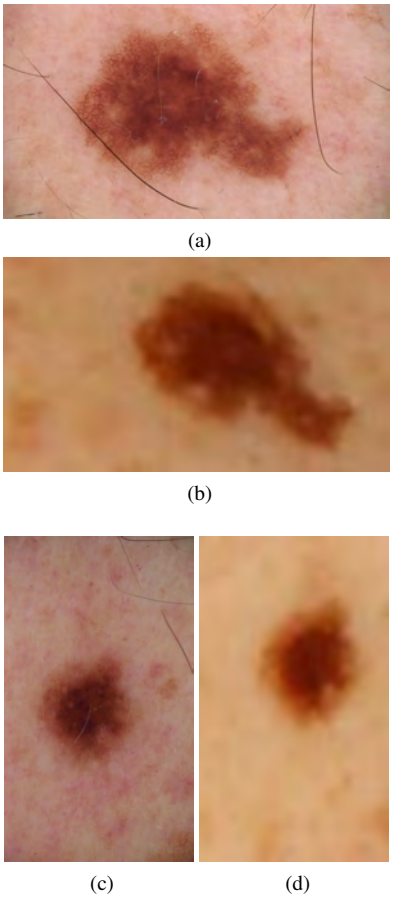
physician's subjectivity. In this context, the idea of developing a total body skin scanning system is a valid alternative to perform this process automatically.

In general, images of skin lesions can be acquired by using different tools (Korotkov and Garcia, 2012). For example, dermoscopy is a non-invasive imaging technique for PSLs that provides the visualisation of their subsurface structures by means of a hand-held incident light magnifying device (microscope). Cross-polarised light is nowadays used, and it allows almost identical images to be obtained using a microscope with or without immersion fluid and direct skin contact with the instrument (Benvenuto-Andrade et al., 2007). Another modality to acquire PSLs consists of dermatological photographs - referred to as clinical or macroscopic images - showing single or multiple skin lesions on the surface of the skin. These images reproduce what a clinician sees with the naked eye (Day and Barbour, 2000). Clinical images are used to document skin lesions, mapping their location in the human body and tracking their changes over time. A comparison by the two modalities is showed in Fig. 1.

For this project, a total body scanning system is used to obtain clinical pictures (Korotkov et al., 2015). It is now at its second version and is capable of acquiring high-resolution images of the skin. Fig. 2 is an example of a photo acquired by the system and moreover Fig. 3 presents a comparison of two different skin lesions detected on the total body skin images and a dermoscopy tool.

A software framework to detect and compare skin lesions has been already developed at the University of Girona, as detailed in Korotkov et al. (2015). However, given the outstanding results obtained by deep learning, in this work we implemented a new workflow based on artificial intelligence, to automatically localise and characterise the lesions and to perform intra-exploration matching of them in the images obtained by a single acquisition session. More in details, the main steps on which this work focuses on are 1) detection, 2) segmentation and 3) intra-exploration matching of the pigmented skin lesions. A scheme of the proposed pipeline is shown in Fig. 4. The first step is related to detect the lesions on the total body photos. In fact, the ac-

Figure 2: Example image acquired by the new total body skin scanning system.



(a) (b) (c) (d)

Figure 3: Comparison between the same PSLs obtained with MoleMax digital dermoscopy (*a* and *c*) and extracted by the skin body image (*b* and *d*).

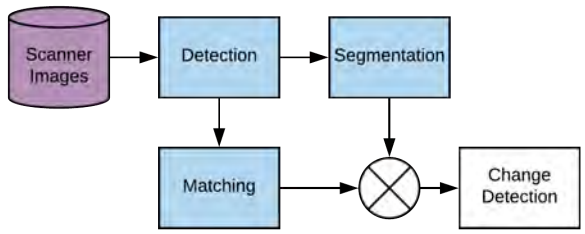


Figure 4: Scheme of the proposed pipeline considered in this project.. Change detection algorithm was not explored but is left for future investigation.

quired pictures cover large parts of the skin surface, in which many PSLs can be found. Because each of the them may change in a future acquisition, the first primary objective is to collect and label the largest number of skin lesions. The second step that has been solved involves the segmentation of lesions on the ROIs localised by the detection technique. As stated by Korotkov and Garcia (2012), border detection (segmentation) of a skin lesion is crucial for its automated diagnosis and is one of the most active areas in the computerised analysis of skin lesions. After their detection and segmentation, the lesions have to be matched to create a unique map. Matching pigmented skin lesions is an important step of the pipeline and its application is required for two different reasons. First of all, because of the architecture of the scanning system, adjacent cameras acquire overlapping regions of the skin surface during a single acquisition. Therefore, the same skin lesions can be found in different images in which they appear with different views according to their position with respect to the acquiring cameras. We want to maintain their multiples views to estimate the 3D orientation of the lesions. But at the same time we need to discard duplicated representations of the same PSLs. To address this, it is important to understand if a skin lesion has its counterparts in other images and, if this is the case, match them in order to refer all the views of a PSL to a unique map. Since it is related to performing a mapping procedure for the PSLs from a single acquisition, this process takes the name of intra-exploration matching. Another reason for which the matching task is required, comes from the change-detection algorithm. In fact, in a follow up for melanoma detection, a patient would be scanned several times. It means that for each acquisition, an intra-exploration map is created. Then, the maps have to be compared with the baseline to look for future changes happened during the inter-explorations period. To control the evolution of each skin lesion, we have to match its versions acquired at different times in order to compare them correctly. However, for the moment, we focus only on the intra-matching task, by leaving the inter-matching as future work.

The rest of the document is organised as follows. In section 2, an analysis of the state of the art for the various steps is presented. In section 3, the dataset available is described. Then, in section 4 the methodology chosen for this project is presented with more details. In section 5, the results of each step are presented together with a discussion about them.

2. State of the art

The proposed review takes in account not only techniques related to clinical images of the skin, but also more techniques of computer vision in general. The main tasks that have been reviewed in this work are three: detection of skin lesions, their segmentation and then intra-examination matching.

2.1. Detection of the Skin Lesions

The detection task includes two different steps that are usually accomplished by the reviewed algorithms (Girshick et al., 2014; Girshick, 2015; Ren et al., 2015). The first one is related to the spatial localization of the objects, that is usually expressed with rectangular bounding boxes around them. Their coordinates are given with respect to the origin (0,0) in the image of interest. In the algorithms that have been analysed, it is common practice that these coordinates are composed of four elements: x and y coordinates of the top left of the bounding box, plus its width and height. The second element is a vector of probabilities related to the classes to which the object belongs.

For what concerns our work, total body photography has started to be investigated recently, and as consequence of this there are few reported applications of object-detection algorithms for skin lesions on full body images. Whilst this lack of information represents an unexplored research field that in our opinion gives particular value to our proposed technique, it also presents a challenge in finding scientific literature on which we could base our methods. For this reason, a review of the state-of-the-art techniques for general detection task is proposed. For years the detection has been based on the use of hand-crafted features, such as histograms of gaussian (HoG) presented by McConnell (1986) etc.), followed then by a trainable classifier, for example, Support Vector Machine - introduced by Cortes and Vapnik (1995) -, boosted classifiers or random forests. Besides, in the pipeline developed by Korotkov et al. (2015) for the previous version of the scanning system they make use of a method widely used for blob detection, the maximally stable extremal regions (MSER) method, presented by Matas et al. (2004). Its application is based on the idea that usually, the PSLs are darker spots on the lighter skin surface. Regarding this, before using this detector, they apply two pre-processing steps to ease detection of the PSLs. The first one is a foreground (skin)

segmentation, which afterwards is used 1) to exclude from the PSLs detection areas not belonging to the patient body (scanner's surfaces in the background etc.) and 2) to selectively enhance the skin lesions w.r.t. the skin surface. From their results, the MSER detector applied on the scanner images yields stable skin lesion regions (blobs) which are approximated by ellipses with their respective dimensions (major and minor axes) and locations (centre coordinates). However, this technique is an intensity-based method that is not able to detect such a large number of skin lesions, since they can include many freckles which are not so easily identified because of their intensity range being very close to the skin of scanned patients.

Furthermore, the design of an algorithm for features' extraction is difficult to engineer without the help of some learning procedure. A considerable improvement has been possible thanks to Convolutional Networks (ConvNets), that are examples of hierarchical systems with end-to-end feature learning that are trained in a supervised fashion. The use of deep learning approaches for object detection represents a new challenge compared to other computer vision tasks (classification, recognition, segmentation, etc.). However, it can be found that the main approaches based on artificial intelligence implement the localisation tasks as slight variant of the classification step. Regarding this, the idea of this work is to replace the previous method with a deep learning approach, that has proven to produce outstanding results in detection tasks[2]. Their effectiveness is clearly visible if we analyse the results of public challenge in the visual recognition task. As an example, in the PASCAL VOC object detection[3], the best results are obtained with convolutional neural networks. However, to the best of our knowledge, there is no evidence of the application of artificial intelligence techniques for detecting PSLs on clinical images. Therefore, in the paragraph below, a review of deep learning - based algorithms applied to the detection of general objects (cars, pedestrians etc.) is presented.

One of the first techniques built with deep learning techniques has been the Region-based Convolutional Neural Network (R-CNN) presented by Girshick et al. (2014). It is based on the combination of Selective Search (Uijlings et al., 2013), a region proposals algorithm, and a convolutional neural network applied for each of the generated ROI. The output is then passed to an SVM, that labels the regional, and a linear regressor that can insert a bounding box around the detected objects. The idea on which R-CNN is based is simple and successful, its only problem is its low speed due to the exhaustive Selective Search technique. Its successive algorithm is an implementation of a fast version of the R-CNN that not by chance is called Fast R-CNN (Girshick, 2015). It has different analogies with its predecessor, but with improvements in the detection speed. There are two main modifications: 1) a feature extraction over the image before proposing candidate regions, 2) the use of a softmax layer instead of SVM. The first change makes use of one CNN over the entire image, instead of multiple ones as in R-CNN, and the second one extends the neural network predictions instead of creating a new model. However, this approach makes still use of Selective Search for regional proposal, the main bottleneck for what concerns the speed. Finally, another improvement has been accomplished in the Faster R-CNN algorithm by Ren et al. (2015) by replacing the slow object proposal method of the previous two approaches with a fast neural network, called regional proposal network (RPN). The most relevant modification of this algorithm resides in the use of a sliding window over the feature map of an initial CNN to map it to lower dimensions. For each location taken into account by the sliding window, multiple possible regions are proposed. They are called anchor boxes, and each of them comes with a bounding box with default dimensions and ratio. Then, the RPN network takes each of these proposals and give them an *objectiness* score (background VS foreground) and bounding boxes coordinates. If an anchor box has an objectiness score above a certain threshold, its box coordinates get accepted as region proposal. After that, this information is feed into a Fast R-CNN, that is identical to the network of the previous approach. In the end, Faster R-CNN technique achieves better speeds than before and state-of-the-art accuracy. Later models did a lot to increase the detection speeds, but few of them managed to outperform Faster R-CNN by a significant margin. In conclusion Faster R-CNN can be considered a canonical implementation for object detection. However, successive algorithms tried to improve the speed of the object detection, and one of them is the Region-based Fully Convolutional Net proposed by Girshick et al. (2016), that is several times faster than Faster R-CNN and achieves comparable accuracy. In fact, it can simultaneously address location variance by proposing different object regions, and location invariance by having each region proposal refer back to the same bank of score maps. Moreover, it is fully convolutional, so all the computation is shared throughout the network. All the previous approaches are based on the use of two separate steps: regional proposal and region classifications. Another pipeline is proposed in Single Shot Detector (SSD), where Liu et al. (2016) combine the two steps in one single shot.

2.2. Segmentation of the Skin Lesions

After the extraction of regions of interest, the PSLs are segmented. As stated by Korotkov and Garcia (2012), one of the earliest works on skin lesion border detection used the concept of spherical coordinates for color space representation (Umbaugh et al., 1989).

[2] http://image-net.org/challenges/LSVRC/2015/results
[3] http://host.robots.ox.ac.uk/pascal/VOC/

Since then, it has been widely adopted in the literature for lesion feature extraction and color segmentation. Comparisons of different color spaces applied to segmentation were carried out by Umbaugh et al. (1992, 1993). Golston et al. (1990) estimated the role of several determinants of the lesion border, namely color, luminance, texture and 3D information. While 3D information was mostly absent, color and luminance appeared to be the major factors for most of the images. Thus, the authors discussed an overall algorithm that would take into account several border determinants based on their level of confidence, and proposed a radial search method based on luminance information. Similarly, in support of multifactorial descriptiveness of the lesion border, Dhawan and Sim (1992) proposed the combination of gray-level intensity and textural information. Further works concentrated on improving existing techniques (Zhang et al., 2000) and applying a multitude of different approaches, including edge detection (Denton et al., 1995; Xu et al., 1999), active contours (Chung and Sapiro, 2000), PDE (Barcelos and Pires, 2009; Chung and Sapiro, 2000), gradient vector flow (Tang, 2009) among many others.

Recently, the community has started to move from traditional techniques towards deep learning techniques, following the general trend of computer vision (Fornaciali et al., 2016). Among the challenges hosted by the International Skin International Collaboration (ISIC) sponsored by the International Society for Digital Imaging of the Skin (ISDIS), there is the segmentation of dermoscopic images challenge. Several approaches that obtained top results are based on the U-net of Ronneberger et al. (2015), a convolutional network intended for accurate segmentation of biomedical images. One of the these algorithms[4] has been implemented by Campinas University in Brazil[5], as described by Menegola et al. (2017). The convolutional neural network has been built specifically for dermoscopic images.

2.3. Matching of the Skin Lesions

Multiple occurrences of the PSLs in various images obtained in a single acquisition of the scanning system have to be matched in order to create a unique map of each skin lesion. A common strategy is to characterise each element by generating a same-size vector that can uniquely describe it. In this way, the matching problem can be seen as a process in which objects with similar descriptors (whose distance is lower than a given threshold) are identified to match. In computer vision, generating descriptors has been a hot topic in the last decades. One of the milestones for descriptors extraction comes from the Scale Invariant Feature (SIFT) method by Lowe (1999). It allows generating spatial-invariant descriptors for objects of interest, to find correspondences useful to the matching process afterwards. This method has also been used by Korotkov et al. (2015), where the skin lesions and larger skin regions are compared using SIFT. Specifically, in their pipeline for matching, the authors initially consider the stereo-pair images for matching and subsequent triangulation of the MSER blobs. In fact, the detected skin lesions are compared using image feature descriptors (SIFT) and then triangulated by using the intrinsic and extrinsic parameters of the stereo rig. Once the PSLs are matched across all the stereo pairs, their 2-D views at each turntable position are grouped in sets. Afterwards these sets are generated for each turntable position and a matching process across different turntable positions is performed with a similar approach of before, based on the comparison of features extracted by SIFT descriptor on ROIs of the detected skin lesions. At the end of this method, every PSL has all its 2-D views in the images of the scanner correctly referring to it.

Several other approaches have been used for matching skin lesions. One of them comes from the work by Perednia and White (1992), in which a correct identification of initial matches works as base for a 3-point geometrical transformation. The same authors proposed a method to automatically extract initial PSLs matches by using a Gabriel graph representation of lesions in an image. The same process is a requirement for the baseline algorithm proposed by Roning and Riech (1998). They first create lesion maps to perform the registration of multiple lesion images. Another approach comes from Mcgregor (1998). Clusters of nevi are generated by using a centre-surround differential operator and later at different images scales they are thinned via a centring mask. A registration process is then performed, but it requires initial lesion matches, which are obtained by minimising the distance and angular error of local neighbourhood. Instead, the authors Huang and Bergstresser (2007) used Voronoi cells to measure similarities between skin lesions. Another approach based on graph matching was proposed by Mirzaalian et al. (2009).

All previous approaches do not include any deep learning techniques. However as it is happening in other computer vision tasks, for what concerns generation of descriptors, convolution neural networks outperform the handcrafted ones, as stated in Fischer et al. (2014). Starting from the success of CNNs trained on public dataset such as ImageNet in recognition tasks obtained by Krizhevsky et al. (2012), they demonstrated how convolutional neural networks outperform SIFT on descriptor matching.

[4] https://challenge.kitware.com/phase/584b0afacad3a51cc66c8e24
[5] https://www.unicamp.br/unicamp/english

Figure 5: A 3-D model of the scannining system: exterion view on the left and acquisition compartment with side door removed (right). Its human-size dimensions allows patients to stand up during the acquisition.

3. Material and Methods

3.1. Total Body Skin Scanning System

The first version of a photogrammetry-based total body scanning system was developed by the ViCOROB institute at the University of Girona in 2015. It is equipped with 21 high-resolution cameras organised in two columns and a turntable, and it allows the acquisition of overlapping images, covering almost 85 - 90 per cent of the patients skin surface. The tests performed with this scanner showed that it could be used for automated mapping and temporal monitoring of multiple lesions. A new version of the scanner has been built by the same team. It has eleven Canon 6D cameras with CMOS sensor of 20.2 megapixels, that can cover the same skin surface as before. The lighting procedure is not continuous as before, but a set of flashes is used to light the scene, and they are synchronised with the shooting time. This modification increases three times the amount of light, allowing better illumination conditions and better light polarisation. Also, lower ISO is used, and consequentially, image quality is increased concerning the previous implementation.

The system appears as a human size cabin (in Fig. 5), and it is equipped with a rotating platform. A patient enters inside and locates his/her feet on the moving stand. The scanning is performed in two steps. Initially, the person has to grab a support in front of him/her and the platform performs a rotation of 180 degrees counterclockwise, with images acquired in twelve fix turntable positions. After the turntable reaches 180 degrees, the patient needs to change his/her position and to grasp the same support but with a posterior pose. Then the same amount of images are acquired. In total, the final amount of images per scanning session are 264. For the moment, the scanning system is located inside a room on the laboratory. In the future, after the technical experiments will be finished, clinical tests will be performed at the Hospital Clinic in Barcelona (Spain), starting a follow up of a larger amount of patients.

3.2. Dataset

The dataset is composed of clinical images - each one with sizes 5472x3648 - coming from the total body skin scanning system. One patient has been scanned in two different sessions, so the final number of images we used for this project is 528. Our pictures are acquired with a larger field of view compared to other acquisition methods (i.e. dermoscopy), so that they may include skin areas with different PSLs in them. For this project, the primary interest is in the photographs in which the skin lesions are well visible. For instance, those covering the back of a patient have been chosen as good candidates: they include flat skin surfaces, in which the PSLs are clearly displayed. On the contrary, the hands and the feet have a shape and other characteristics (as illumination conditions) that do not allow to observe very clearly the skin lesions, so for this reasons they have been discarded. Furthermore, the head skin has been excluded, because possibly covered by hair and because of the different curves of the various elements forming the face, that does not let the skin lesions to be seen.

Unfortunately, we were not able to get a ground truth from a trained physician during the course of this work. As consequence, a qualitative analysis has been accomplished to verify the results of detection and segmentation tasks in the implemented pipeline. Moreover, original arrangements - as we explain in the next section - have been taken to produce the training, validation and test sets to run the CNNs for the various steps. Then, specific modifications and augmentations to the original dataset have been performed for each task, i.e. 1) detection, 2) segmentation and finally 3) intra-acquistion matching of the skin lesions. For instance, in the segmentation part, a public dataset of dermoscopic images has been used. This dataset comes from ISIC Challenge 2017 "Skin Lesion Analysis Towards Melanoma Detection" (ISBI 2017).

4. Detection and Characterisation of Skin Lesions

To detect, characterise and match the skin lesions contained in the images acquired during one acquisition of the scanning system, we propose a pipeline organised with 1) an initial detection of the PSLs, followed by 2) their characterisation based on a segmentation approach. After these two steps, 3) the intra-exploration match of the skin lesions is accomplished by using the localisation provided by the detection algorithm. State of the art deep learning techniques have been used to solve the intended tasks, and a summary of our work is presented below.

4.1. Detection of the Skin Lesions

For the first task, we decided to use Faster R-CNN, which reached state-of-art in public competitions. We take advantage of the public implementation of the network available at the Facebook Detectron Project[6].

As backbone of convolution body of the detection algorithm, VGG16 by Simonyan and Zisserman (2014) has been replaced with Deep Residual Network (ResNet with 50 layers). In fact, as stated by He et al. (2016) and Ren et al. (2015), the use of deep residual learning allows having broader networks - able to learn a wider range of features (Szegedy et al., 2015) - that are more easily trainable and at the same time they can gain accuracy because of the increased depth. Furthermore, the connections based on Feature Pyramid Network (FPN) have been added to the ResNet. It has been showed by Lin et al. (2017) how the inherent multi-scale, pyramidal hierarchy of deep convolutional networks are able 1) to build feature pyramids with a marginal extra cost and 2) to obtain better results in Faster R-CNN. Besides, FPN has recently enabled new top results in all tracks of the COCO competition, including detection, instance segmentation, and keypoint estimationee (He et al., 2017).

As is common practice (Girshick et al., 2014), a pretrained ResNet-50 model is obtained from the MSRA (Microsoft Research Asia) repository, which contains also the models of the Deep Residual Network. Fine tuning, based on slightly updating the weights of all the level of the Faster R-CNN, has been used to train the network. To perform this task, it has been necessary to create a specific dataset, that includes 1) the images in which the PSLs have to be detected and 2) the annotations (in JSON format) of the skin lesions in the photos. The annotations file is composed of bounding box coordinates per object, specified as described in subsection 2.1 of the state of the art paragraph. The images of our dataset come with no annotations, so these have been created for this specific purpose. There are software programs[7] that can be used to manually annotate the files of interest. However, in our particular problem, the process cannot be performed manually, because the number of skin lesions is considerable, and the required time to do the annotations would have been unfeasible. Instead, it was decided to take advantage of the MSER detection - segmentation method used by Korotkov et al. (2015), followed by a boundary detection[8] and minimum bounding box detection. This method allowed us to generate boxes for a large number of PSLs in the images of interest. This ground truth generation method is based on the good results in PSLs detection obtained by Korotkov et al. (2015): starting from them, we generated our dataset to feed the Faster R-CNN.

Another modification that has been done is related to the size of the images processed by the CNN. The good results of Ren et al. (2015) are obtained for COCO dataset, whose images have sizes of one order of magnitude smaller compared to the one in our dataset. For this reasons, the initial attempts of the experiments were not able to obtain the desired output, even changing the available parameters of the code. To resolve this, each image has been divided into smaller overlapping patches (one hundred per each picture), with a dimension (484×668) similar to the ones of COCO dataset.

Furthermore, several parameters have been tuned in the Regional Proposal Network. The modifications that have been done take into account the characteristic of the object we are willing to detect. First, the default anchor ratios were set to 0.5, 1, 2. However, it is not likely for the PSLs to be enclosed in a rectangular box with one side dimension double of the other. For this reason, they have been changed to 0.7, 1, 1.5. Another consideration is done by taking into account the dimensions of the skin lesions in the patches. The default anchor sizes are set to (64, 128, 256, 512). However, in our dataset, it has been observed that it is difficult to have a PSL larger than 120 pixels per side. So, the dimensions of the anchors have been reduced to concentrate only on smaller objects (32, 64, 128), and also the possible size configurations have been fixed from five to three, because of the absence of so much variety in dimensions of the objects to detect. This task ends with the detection of skin lesions by drawing bounding boxes around them. For our pipeline, it has been decided to use larger windows than the predefined ones to save the detected PSLs. This decision involves consequentially to include wider parts of the skin surrounding the lesions, to allow a better characterisation in the following steps.

4.2. Segmentation of Skin Lesions

The next step of the pipeline is the characterisation of these region of interest. To do that, it has been decided to segment the border of the skin lesions. This task has been solved by using the implementation of the U-Net by Menegola et al. (2017), which reached outstanding results in the segmentation task on ISIC challenge[9]. The challenge includes only dermoscopic images, on which Menegola et al. (2017) fine-tuned their neural network. Following a similar approach, several experiments have been conducted to fine tune the network, by using different types of datasets. First, the training has been built only with the 2000 dermoscopic images available for training on the challenge from ISIC. Then, a mixed training set has been built by combining the dermoscopic images of the challenge with the detected skin lesions coming from the detection tasks. To obtain the ground truth masks for such images, the bounding box

[6] https://github.com/facebookresearch/Detectron

[7] https://github.com/tzutalin/labelImg

[8] https://www.mathworks.com/help/images/ref/bwboundaries.html

[9] https://challenge.kitware.com/phase/5841916ccad3a51cc66c8db0

coordinates associated to each detected skin lesion are used to clip also the detection maps obtained by MSER algorithm. As last test, we train only on the images of our dataset. In the final solution, we have been decided to train the neural network by mixing the two datasets in equal percentage. Regarding this, even they come from another acquisition modality, the dermoscopy images have been already used in combination of clinical images (as an example, we cite the classification task of Menegola et al. (2017)). For this reason, we took advantage of the well-done masks of the dermoscopy images. At the same time, we are sure to have included the characteristics of our dataset in the training process. The images are resized to 128 by 128 in the default configuration. However, it has been seen that the best sizes for having good results on our dataset are 64 by 64. Another modification has been reserved for the introduction of early stopping technique. In the original implementation, the number of epochs has been set to 220. However, with early stopping technique, the results of the original network on the challenge dataset have been found to be the same that the one reported, but with a much shorter computational time (less than 10 epochs).

4.3. Matching of Skin Lesions

After the PSLs have been detected and segmented, they are ready to be matched. In this work, we focus on the matching of the skin lesions to create a map of a single acquisition performed by the scanner. The change detection could be investigated with a similar method to the proposed one, but it has been left as future work. In our approach, we establish a new workflow organised with a first rough registration process performed with a method based on hand-crafted features and geometrical constraints of the scanning system, followed by a novel deep learning technique to refine the matches.

In the proposed pipeline, a first registration method is used to generate rough correspondences in the various images. In particular, after an initial calibration of the intrinsic and extrinsic parameters of the cameras, SIFT features are extracted for the images acquired by the eleven cameras at the same turntable position. Next, the 2-D extracted points are matched together and triangulated to 3-D space. This procedure is performed at each known turntable position, and a 3-D point cloud is obtained by grouping all together the triangulated points at different turntable positions. Then, the 3-D point set is converted to a surface mesh by using the method of Kazhdan and Hoppe (2013). Given the generated 3-D mesh and the known geometry of the system, we are able to infer for a given point in an image, its 3D position on the surface, and where it back-projects in the other images. This technique works well, but the results showed that it represents only a rough estimation of the correspondences between the different areas of the body. The algorithm itself does not give an exact estimate location, mainly because it is very likely that during the acquisition the shape of a patient is prone to involuntary movements that do not allow for an exact matching among skin lesions. Nevertheless, we want to be sure of the locations of the same PSLs in the images, so a refinement process on the results obtained by the previous method is performed.

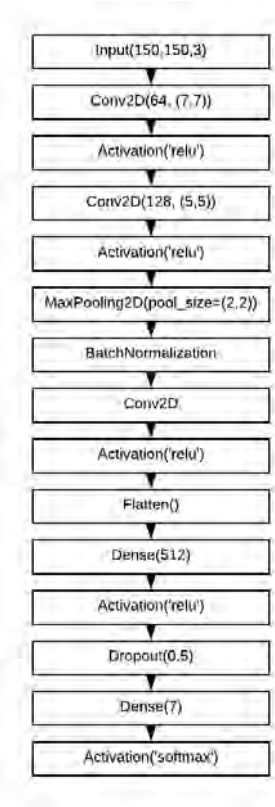


Figure 6: Architecture of the Classification Network created for matching skin lesions

To perform this, a local matching on the areas proposed by the previous technique is performed. The idea is to compare locally the PSLs to be matched and the skin lesions detected in the suggested regions. For this task, it has been decided to take advantage of a methodology inspired by the one described by Fischer et al. (2014). Here, they demonstrate how features extraction by neural networks outperforms the traditional SIFT descriptor for matching tasks. However, we use a slight variant of their network, organised as shown in Fig. 6. To build the training set, eighty-four different skin lesions are used to generate the training and validation set, with a similar method described in the original implementation. Each skin lesion goes under four different geometrical transformations: one rotation of thirty degrees and three different affine transformations. Despite what described in their paper, we decided to generate modifications that resemble the most the different views in which the PSLs can be found in the images, without causing any very variating changes. After that, the neural network is trained with the 336 pictures to classify

them in 84 classes. The results are checked by using a validation set composed of 84 images. All the images are resized to 100×100. The training is conducted for 20 epochs, with an Adam optimiser, a learning rate of 0.0001 and a batch size of 4. After the end of this process, features have been extracted. Descriptors at two different network depths - corresponding to the last two fully connected layers (or dense layers[10]) - have been compared. The weights of the first fully connected layer (fcl) are 110,166,016 plus 512 biases, on the contrary the ones of the second fcl are 43016 plus 84 biases. No feature reduction method has been applied on the extracted descriptors.

5. Results and Discussion

In this section, results of the new pipeline - composed of detection, segmentation and intra-exploration matching of the PSLs - are presented. Several examples are displayed to highlight the improvements of our method with respect to the former one. We recall that no ground truth has been generated with the help of a physician, consequentially we decided to avoid quantitative results for detection and segmentation tasks because they would not have any scientific relevance. Instead, our efforts are focused on presenting visual examples in which our pipeline is able to detect and segment PSLs on the full-body images correctly. Instead, for what concerns the matching process, our methodology allows to generate numerical results.

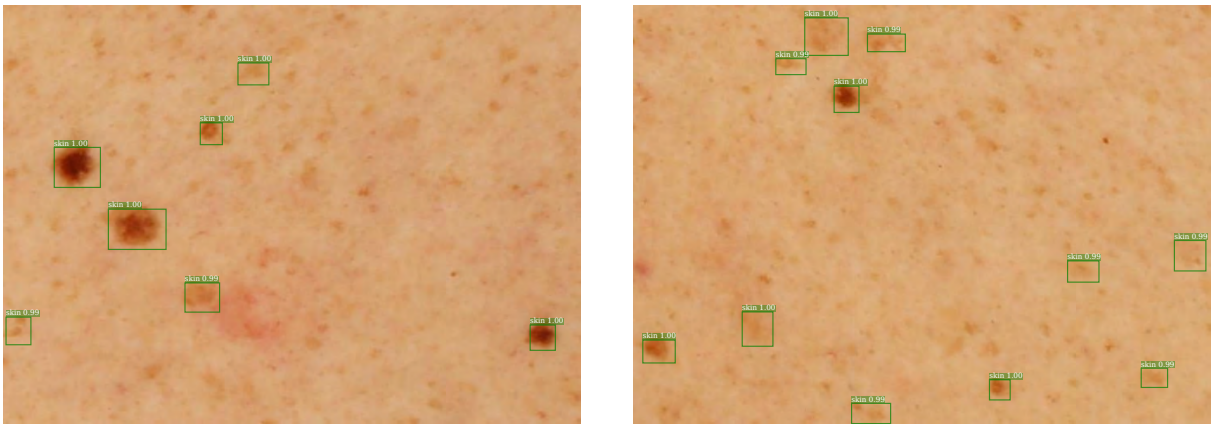
5.1. Detection and Segmentation of Skin Lesions

The detection task has been solved, as Fig. 7 shows. According to our knowledge, the proposed method represents an innovative step in the field of total body skin examination, where no use of deep learning techniques for PSLs detection has been considered yet. We can see how the most evident nevi are recognised by drawing bounding boxes around them, with very high probabilies, even if many lesions detected are freckles. Since in this work no distinction between nevi and freckles is considered, the results are consistent with the fact that we are interested in detecting as many PSLs as possible. In addition, a comparison between our method and the previous one can be performed. Regarding this, because the results of MSER detector are given as PSLs including the mask of their borders, we want to be consistent in the comparison between the outputs of two methods. For this reasons, for Faster R-CNN results we consider the detected skin lesions with the masks generated by the segmentation task. However, in this first part, our focus will remain on the detection task and no observations will be made regarding the different shapes of the PSLs. Regarding this, a comparison between the two methods is showed in Fig. 8, obtained by applying the pre-processing performed by the previous pipeline before MSER detection. On these images, an enhancement of the PSLs is achieved, so that it is easier to highlight them on the skin. They show which of the same PSLs have been detected by the two different techniques. In Fig. 8.*a*, *b* we can underline that both approaches segment correctly the nevi, but as expected the MSER method skips several freckles that have an intensity very close to the intensity range of the patient skin (Fig. 8.*c*-*f*).

Then, the inference process has been applied also to types of images that have not been used in training. In particular, Fig. 9 show the network responding to hairy parts of the body (in our case, they belong to the chest of the patient). In Fig. 9.*a* and 9.*c*, it is evident that the network correctly detects the more clear PSLs, even if they are covered by the black hair of the subject. However Fig. 9.*c* illustrates how the presence of hair prevents the network to detect correctly all the skin lesions (in particular a large number of freckles), by generating several false negative detections. On the other hand it happens also that, because of the conformation of the hairy parts, several of these regions are detected wrongly as skin lesions (Fig. 9.*e* shows an example of false positive). However, it can be observed how the performance of the new approach is improved with respect to the previous method, even if it is still inevitable that in the clinical setting, the patient will be asked to shave the hairy parts. In fact, Fig. 9.*b*, 9.*d* and 9.*f* show that MSER algorithm generates more false positive as well as false negative detections. This is due to 1) the pre-processing applied to the image to enhance darker spots, that is not able to selectively discriminate between skin lesions and hair of the patient, and because the 2) the technique is not sensible to the different of textures of these kind of areas. The MSER is mainly an intensity-based method, but as in this case the hairy regions have an intensity range close to the darker skin lesions, it is not able to differentiate them correctly. On the other hand, Faster R-CNN is a deep learning approach that perform its detection based on a richer collection of features, that improve (but not solve) the detection process. Moreover, we tested the network on parts that contain dark undergarment worn by the patient. Fig. 10 generated by the combination of detection and segmentation show no false positive in these areas. It is in accordance with the previous technique, that also obtains good results with a dark undergarment worn by the patient.

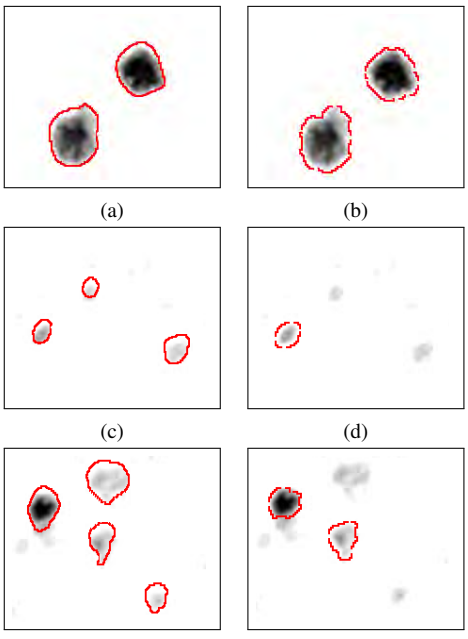
For the second task, each detected skin lesion has been segmented, and Fig. 11 illustrates some examples. In particular, we can see how the borders of both nevi and freckles (Fig. *a* and *b* and freckle in *c*) agree correctly with the PSLs borders. For what concerns the hairy areas detected by the Faster R-CNN, Fig. 12 present the segmentation output on them. In Fig. 12.*d* we can see that, since the hair presence generates

[10] https://keras.io/layers/core/dense



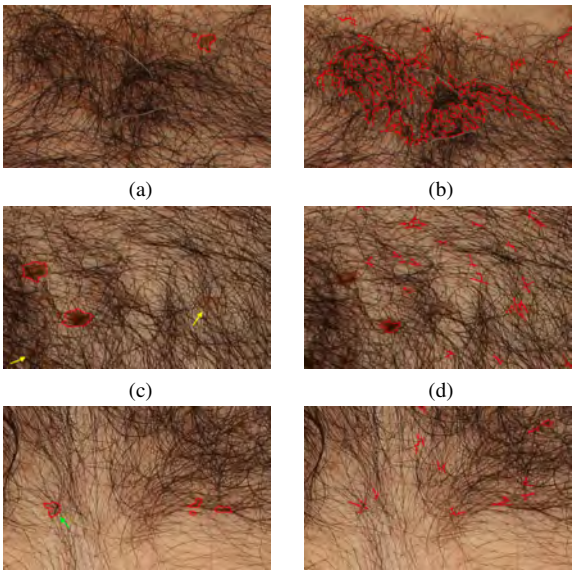
(a) (b)

Figure 7: Output of skin lesion detection with Faster R-CNN on a region of the back of the patient. No classification between nevi and freckles has been performed, because for the moment we are interested in detecting as many PSLs as possible.



(a) (b) (c) (d) (e) (f)

Figure 8: Comparison of the two detection algorithms by overlapping the detected skin lesions on the pre-processed images that are demanded before MSER algorithm. From them, it is easy to see how the freckles are the cause of different false negative in the previous approach, because of their intensity very similar to the skin of the patient. On the contrary, Faster R-CNN bases its detection on the enriched features learnt during the training, and its false positive rate is lower compared to MSER.)

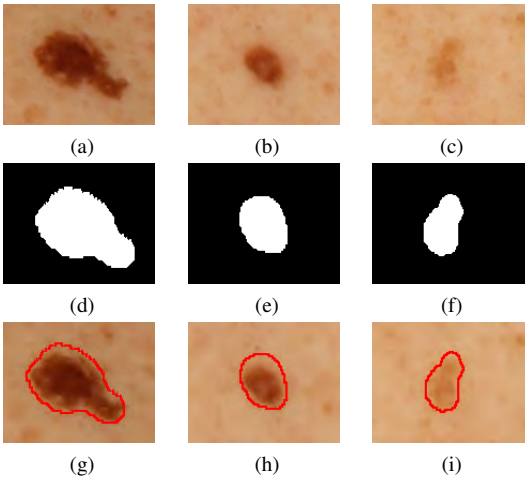


(a) (b) (c) (d) (e) (f)

Figure 9: Detection results for hairy areas. The two algorithms are compared (a, c, e by the new approach and in b, d, f by MSER). It is evident how the MSER approach generates more false positive with respect to our method, that instead can correctly most of the nevi. However, in c we can notice how in the hairy regions false negative (yellow arrows) are present, as well as in e a false positive is highlighted by green arrow. Even if the number of wrong detections is reduced compared to the previous method, the indication of shaving these kind of area is still a recommendation.

(a) (b)

Figure 10: Example of detection of PSLs on skin area in which the patient wears dark undergarment. Both the approaches perform well, with no false positive detected on the trouser. In Fig. *a* detection performed by Faster R-CNN, in Fig. *b* detection performed by MSER approach.

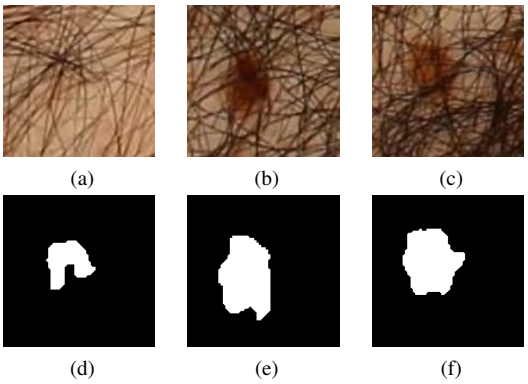


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

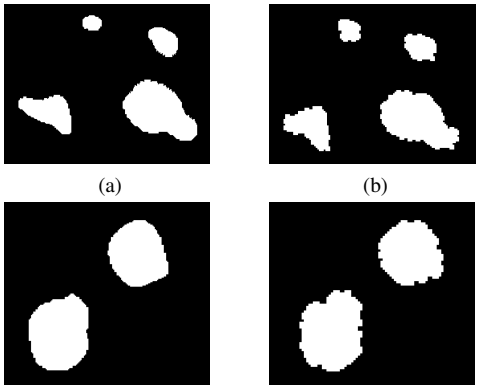
Figure 11: Output of the segmentation by the U-Net for two different PSLs (nevi in *a*, *b* and freckle in *c*). It is evident to see how the the generated masks (shown in *d*, *e* and *f* figures) is able to correctly segment the borders of the PSLs (*g*, *h* and *i* figures)



(a) (b) (c)

(d) (e) (f)

Figure 12: Segmentation results on areas detected by Faster R-CNN on hairy part of the body.



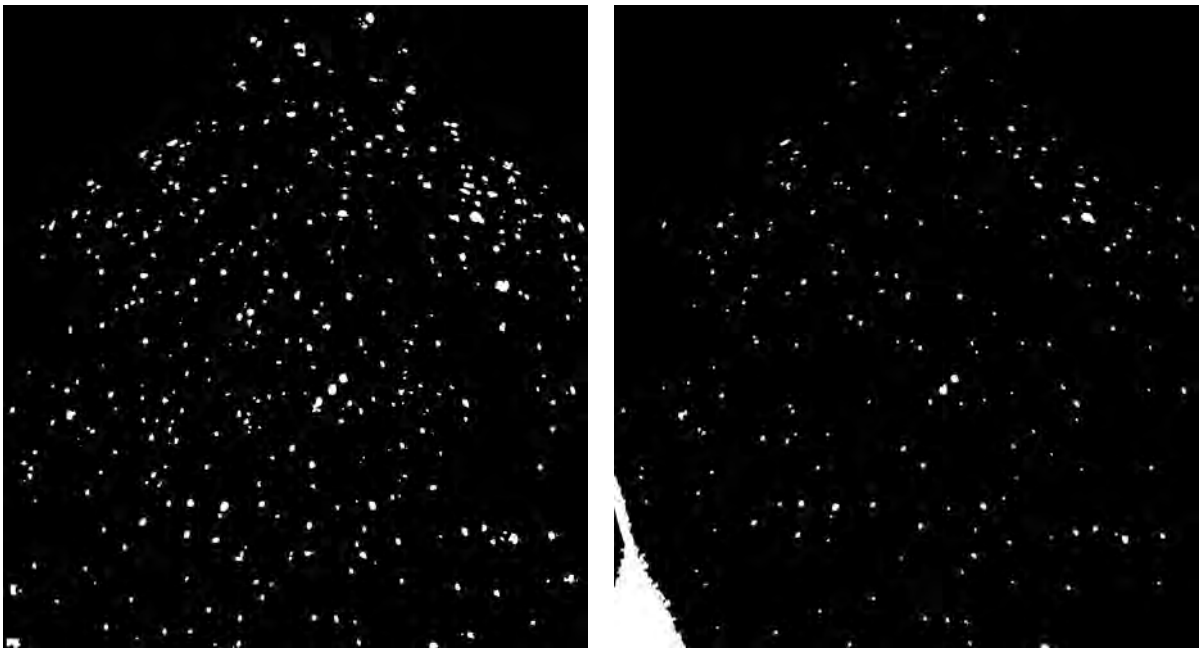
(a) (b)

(c) (d)

Figure 13: Segmentation comparison of the two methods (U-Net results in *a* and *c*, MSER result in *b* and *d*). It is evident how the our method agree with the previous one on border detection.

false positive in detection, the segmentation algorithm wrongly produces a mask on a hairy region. Moreover, Figs. 12.*e* and 12.*f* show that the segmentation is accomplished, but with no precision in the border detection. These wrong results are again due to the intensity similarities between darker skin lesions and hair. Then, a comparison between our approach and the previous one in the segmentation of some PSLs is presented also in Fig. 13. In particular, it can be seen how the borders of most of the lesions obtained by the latter follow quite well the former. Besides, we can highlight how the profiles of borders of the PSLs processed by the new approach are more realistic and more smoothed.

By combining the detection and segmentation on all the regions of an entire image obtained with the scanning system, a complete mask with the detected PSLs is obtained, as Fig. 14.*a* shows. In Fig. 15 the generated mask is used to highlight the nevi on the original image. This accomplishment of the pipeline is comparable with the previous method, whose segmentation is given by Fig. 14.*b*. In particular, the first thing that can be stated is that in the previous approach the application of MSER algorithm requires two pre-processing



(a) (b)

Figure 14: Comparison of the whole mask segmentation generated by the two methods for the same image obtained with the scanning system. Fig. *a*: Complete Mask generated by our method. After having detected and segmented the PSLs in area extracted by the whole photo, our pipeline is able to generate a complete mask of the initial image. The number of PSLs detected is larger than the other method in Fig. *b*. Fig. *b*: Complete Mask generated by MSER method. This is segmentation should be compared with the one performed by our new method. It is evident how the number of detected skin lesions in the former is smaller the latter. Moreover, in the bottom left corner of the image it appears an error in the segmentation due to shaded underlit area.

Figure 15: Overlap between segmentation whole mask and original image. It is evident how the majority of the skin lesions and freckles are detected and segmented.

steps on the images. The first one is the foreground (i.e. skin) detection, to avoid the algorithm to detect skin lesions on the background given by the cabin of the scanner. Another pre-processing resides in the intensity's enhancement of the skin lesions so that the MSER algorithm can perform better in blob detection. On the contrary, our approach is a homogeneous combination of two steps (detection followed by segmentation) that does not require any pre-processing. Moreover, another drawback of the previous approach resides in the false positive obtained by MSER algorithm in shaded underlit areas, such as in the lower-left corner in Fig. 14.*b*. Our algorithm does not present this problem, as Fig. 15 shows. Then, a visual comparison of the same image processed by the two algorithms can be analysed in Fig. 16. It shows where the two segmentation methods agree (yellow color), plus the segmentation performed by the MSER approach (green color) and by our pipeline (red color). From the previous image, it is possible to notice that the latter approach is able to segment most of the skin lesions detected also by the former one. Besides, our technique is able to localise a larger number of nevi. Remember that the first requirement that we decided to accomplish in the beginning was to detect the largest number of skin lesions as possible because each of them could lead to a possible melanoma if a change in its characteristic happens in a future acquisition. So, as image shows, our method

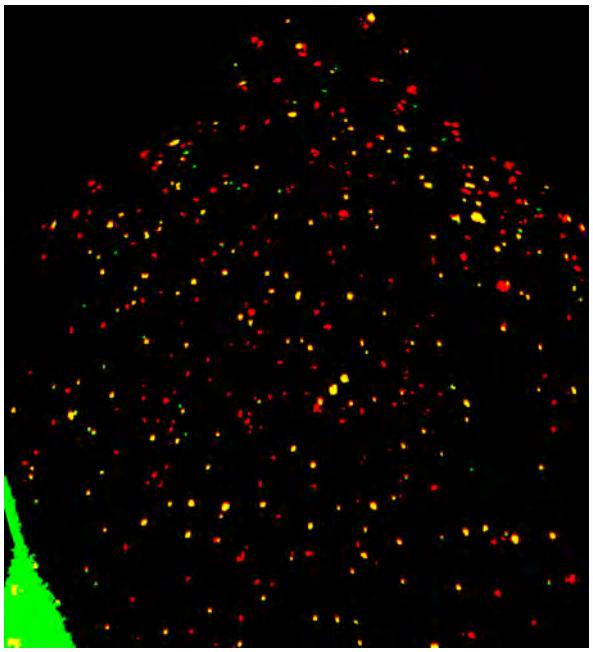
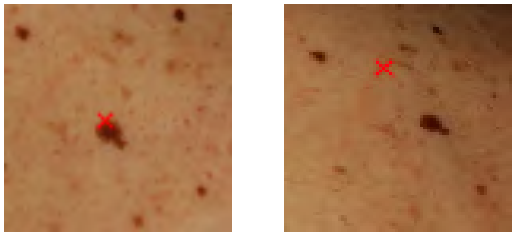


Figure 16: Differences between the segmentation of new algorithm (red color) and MSER method (green color). The yellow color identifies the overlapping PSLs segmented by the two approaches



(a) (b)

Figure 17: Example of a reference point in Fig. *a* for which a matching proposal in Fig. *b* is computed in another image. Even if the proposed region is close to the desired one, it is not precise. For this reason, a further step of refinement is needed.

obtains better results in collecting skin lesions.

5.2. Matching of Skin Lesions

The detection of the PSLs in the different images of the full-body skin of a patient also poses the need of matching them. The initial rough estimation of the locations between a reference point in one image and a matching localisation in other images is not a precise procedure, as the Fig. 17 shows. For this reason, a refinement process finds its use to match correctly the PSLs of interest. The neural network for features extraction has been tested by comparing the descriptors from the last two fully connected layers. The tables presented below show the results in term of L-2 norm and they show which skin lesions in different images (upper row, with a vertical line dividing different images) match the specific PSLs in a given image (column on the left side). Moreover, the tables give a measure of how accurate a match is. In particular, we use the ratio by Lowe (1999) (we refer to it as Lowe's ratio) between the first closest match and the second closest one and if its value is greater than 0.80, the match is not considered good.

As illustrated by the first four rows of Table 1, the features extracted from the two layers allow obtaining vectors that characterise correctly the PSLs in different images from cameras at the same turntable position. In fact, the smallest values in term of Euclidean norm are obtained by the correct matches of the skin lesions between different images. We check if they are good candidates by using the ratio between the closest and second-closest matches. For the match of the first skin lesions to the first image of camera 4, the ratio obtained from the second fully connected layer is 0.31, compared to the 0.44 obtained from the first dense layer. Then, the related values for the first skin lesion with the second images are 0.60 and 0.65. For the match of the second skin lesion, we obtain 0.44 for the first match obtained with the second fully connected layer, compared to the same value of 0.44 obtained from feature vector extracted from the first fully connected layer. Again, for the second match of the same skin lesions, the related values are both equal 0.43. From these results, it is evident how the matches obtained are right in terms of relative distances, because they are smaller than the threshold of 0.80. The results from the second fully connected layer are slightly better. This may seem unexpected if we consider that the first fully connected layer would allow a better characterisation of the image because of its wider length. However, evidence shows that the best results are obtained with the features from the second fully connected layer that holds less weights, which seem to better describe the lesions. Then, more results of matching PSLs at same turntable positions are given by extracting the features from the first fully connected layer, to further explore its performances. The lower part of Table 1 and the second results on Table 2 reveal that the descriptors are suitable to perform the match between skin lesions in different images at the same turntable position. However, the second row of the Table 2 presents the case in which there is no real match to perform between a PSL and the ones in other images. This situation is unlikely to happen but it is due to a failure of the localisation given by the initial rough estimation. The L-2 norm value of 115 indicates there is a match even if it is not the correct one and its value is in line with the correct matches of the other cases. Moreover, the check to understand if it is a good match gives also a result of 0.67, that is higher than the previous ones but still within the acceptance threshold of 0.8. However, this can be avoided by performing a better initial registration. Another observation is done for the match of the last lesions of table 2 with camera 5. The lowest L-2 norm is given by the correct match, but Lowe's ratio is higher than the threshold. Nevertheless,

Table 1: L-2 norm distances between the features extracted at two different depths of the neural network. For five skin lesions detected in an image of camera 3, matches have been searched among the proposed areas in the images of adjacent cameras 4 and 5. The turntable position is fixed. Rows 1-4 compare the results of the matching for features extracted at two different fully connected layers of the neural network: lines 1,2 (in yellow) show the results for descriptor taken at the 2nd fully connected layer (fcl), lines 3,4 (in white) show the relative distances for the same PSLs described by features extracted from the 1st fcl. Rows 5-7 shows the results of the matching for skin lesions characterised by descriptors obtained from the 1st fcl. For all the cases, their Lowe's ratio values are below the threshold of 0.80, therefore they can consider good matching candidates.

	Camera 4							**Camera 5**						
							Lowe's ratio							Lowe's ratio
	102.43	**16.07**	84.29	52.06	110.29	98.09	0.31	107.08	77.00	**29.57**	49.17	103.58	83.39	0.60
	74.42	87.61	**23.74**	53.21	61.30	59.71	0.44	65.76	**21.71**	66.74	54.29	49.59	49.87	0.43
	284.26	**51.33**	181.57	117.88	244.43	213.07	0.43	268.48	164.73	**74.30**	114.23	231.14	188.76	0.65
	205.42	180.66	**55.12**	123.38	153.16	155.29	0.44	187.98	**53.22**	143.80	121.17	136.70	135.24	0.43
	**127.83**	297.53	199.50	225.57	203.78	215.18	0.64	**88.01**	203.92	259.89	212.33	188.65	215.03	0.46
	237.48	143.32	132.28	**43.39**	158.91	116.21	0.37	204.10	118.98	116.48	**52.39**	147.49	97.26	0.53
	236.58	229.01	164.14	130.65	94.50	**22.42**	0.23	194.55	159.55	193.90	136.90	101.53	**52.81**	0.52

we can conclude that, from the majority of the experiments we conducted, the matching for skin lesions at the same turntable position is well solved by our method.

Next, the matching of skin lesions from the same cameras but at different table positions is performed. Table 4 and 3 compare again the L-2 norm values obtained for different skin lesions by extracting features from the two last fully connected layers. In particular, Table 4 shows the results of the matching for the skin lesions on images acquired in previous turntable positions with respect to the reference to which the match is performed. On the contrary Table 3 presents the output of the process for the positions of the turntable that come after the reference. As it can be observed, the results remain good in terms of L-2 norm and Lowe's ratio for all the first adjacent positions with respect to the reference. However, the more the images are acquired in further turntable locations, the more their performance falls down. In particular, we can notice that the L-2 norm values in rows 1,3 still indicate the right candidates, but Lowe's ratio warns that the chosen matches could be wrong. These results are expected because the task of matching skin lesions from the same camera at different turntable positions is intrinsically more difficult compared to the process conducted among different cameras at the same turntable position. In fact, during the acquisition, only the part of the body placed in front of the flash is lighted, and the side areas of the body present a lower illumination. It normally happens, as in the examples shown in the table, that the skin lesions are detected not only in the illuminated areas but also in the darker surfaces. For these reasons, in the process of matching, the same skin PSLs can be found with different lighting conditions and views, which makes their matching challenging. Comparing the results performed in the implementation by Korotkov et al. (2015), they limit the results of merging skin lesions to the first two adjacent turntable positions. Our method can be considered in line with the previous method because the results remain acceptable until the third closest turntable position.

6. Conclusions and Future Work

The initial tests performed for the detection and segmentation tasks show visually how the implemented methods obtain good results. In fact, not only the most significant nevi but also the freckles are detected and segmented, with visible improvements concerning the previous methods. In particular, the deep learning technique for the detection of PSLs on full body skin lesions represents a novel method in total body scanning technique that, to the best of our knowledge, has not been explored yet. MSER detection is an intensity-based technique, and it requires more pre-processing to

Table 2: L-2 norm distances between the features extracted at two different depths of the neural network. For 3 skin lesions detected in other areas of the image from camera 3, matches have been searched among the proposed areas in the images of adjacent cameras 4 and 5. The turntable position is fixed. All rows show results of the matching for features extracted at the 1st fully connected layer. Row 2 shows a particular case in which the proposed area in camera 5 is wrong. Even if the candidate is wrong, the Lowe's ratio is still within the threshold: this can be considered a failure of the pipeline, but it can be fixed by better refining the proposing region algorithm.

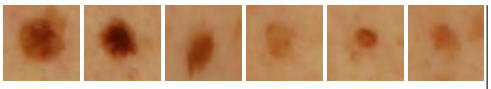
	Camera 4							**Camera 5**							
							Lowe's ratio								Lowe's ratio
	**18.75**	134.68	212.68	290.07	249.72	263.92	0.13	**121.41**	210.67	260.11	197.06	282.41	281.65	249.18	281.12	0.61
	132.07	**24.75**	191.33	275.33	221.09	234.77	0.18	***119.79***	195.02	240.98	177.40	250.19	250.49	237.69	252.37	0.67
	285.76	266.83	187.33	**80.36**	141.11	108.94	0.73	213.72	152.60	214.35	191.44	173.53	148.52	**75.02**	231.64	0.50
	247.42	220.89	124.15	137.51	**41.75**	105.01	0.39	162.36	**90.20**	153.76	124.02	108.88	95.85	111.09	161.09	0.94

Table 3: L-2 norm distances of skin lesions in a specific area of the image of camera 3 at position 17 w.r.t. the PSLs found in the proposed areas of the same camera at different turntable positions. The features are extracted at two different depth: in yellow, the matching performed for vectors taken at 2nd fully connected layer (fcl) is showed, in white, the matching with features extracted at 1st fcl is performed. For the first adjacent position, the matching works in both the cases. On the contrary, more the positions get far from the reference, more errors happen. For instance, in the position 14, even if L-2 norms is good for the same lesion at row 1 and 3, all Lowe's ratios are above the threshold of 0.80.

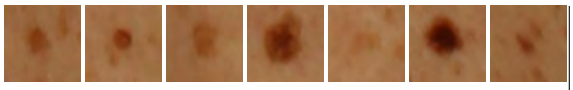
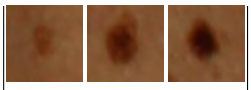
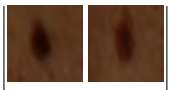
	Position 16								**Position 15**				**Position 14**			
								Lowe's ratio				Lowe's ratio			Lowe's ratio
	106.36	102.40	106.86	**47.48**	119.42	61.25	98.76	0.77	90.70	**57.83**	69.00	0.83	89.90	83.39	0.93
	47.29	**33.83**	43.46	81.75	45.43	99.51	48.69	0.77	**60.45**	85.63	85.55	0.70	75.10	**70.61**	0.93
	232.40	229.90	232.58	**113.34**	281.84	156.34	225.44	0.72	216.48	**151.94**	166.80	0.91	210.09	202.07	0.96
	113.28	**77.88**	96.29	190.84	109.17	219.75	112.26	0.80	**132.25**	194.33	195.54	0.68	182.73	**169.97**	0.92

Table 4: L-2 norm distances of skin lesions in a specific area of the image of camera 3 at position 17 w.r.t. the PSLs found in the same proposed areas of the same camera, but at others turntable positions. The features are extracted at two different depth: in yellow, the matching performed for vectors taken at 2nd fully connected layer (fcl) is showed, in white, the matching with features at 1st fcl is performed. For the first adjacent position, the matching works in both the cases. On the contrary, more the positions get far from the reference, more errors happen. For instance, in the position 20, all Lowe's ratios are above the threshold of 0.80.

	Position 18										**Position 20**				
										Lowe's ratio					Lowe's ratio
	117.04	121.63	74.43	**48.76**	108.61	119.72	114.29	116.52	120.85	0.65	**55.08**	87.08	95.03	55.73	0.95
	36.04	49.68	64.90	83.62	**13.49**	76.70	40.52	38.56	39.59	0.37	86.69	55.21	**52.34**	88.69	0.94
	288.89	294.44	168.06	**106.98**	236.31	303.53	285.64	277.13	297.86	0.63	**142.78**	212.71	217.18	142.88	0.99
	124.41	141.99	177.80	218.39	**33.60**	230.21	128.40	109.78	115.91	0.28	185.64	121.40	**116.90**	196.88	0.95

obtain good results. Therefore, our new methods represent a valid alternative to the previous implementation for the detection task. Moreover, the tests performed for the segmentation task show that the detected skin lesions are correctly segmented, also in comparison with the previous approach. However, to prove the pros and cons of the two strategies, a more thorough and quantitative analysis should be performed. As future work, a physician should indicate with skin lesions are worth to be detected and to provide segmentation on clinical images of PSLs. Then, a classifier will be trained to distinguish relevant PSLs from the rest.

For the matching procedure, the numerical results show how the method performs correctly for PSLs at the same turntable position. In the future, a comparison by using pre-trained neural network will be explored to check how different feature extraction methods work for the same task. Particular attention should be put on matching skin lesions between images acquired at the different turntable positions by the same camera. The results regarding this point show how the neural network allows a good matching, even if with lower robustness. In particular, for the moment, the initial rough estimation of the locations of the skin lesions in other images is not as precise as expected, and this generated several wrong matches because the proposed areas are not the expected ones.

A complete integration of the different steps is still missing. As future work, with the idea of providing a stand-alone software for the full body skin scanning system, all the individual tasks should be integrated together. For the moment, this paper shows how well the different methods work.

7. Acknowledgments

First and foremost, I would like to thank my supervisors Dr Rafael Garcia, Dr Josep Quintana and Dr Ricard Campos, for having given me the opportunity to work under their guidance and to learn from their experiences. They always supported and encouraged me during my work and I am really glad of the final result that we have accomplished. Thanks to you, I obtained that self-security I always looked for, and this is one of the most important thing I will bring with me.
I would like to thank all the professors and assistant professors that I met during the MAIA Master, from the University of Burgundy, from the University of Cassino and Southern Lazio and the University of Girona. You gave me a knowledge I will bring always with me with gratitude. I would like to thanks all the staff of the three universities, for having been helpful to me and the other students and for supporting this valuable and incredible master program.
A big good luck to my friends of the MAIA master, first promotion of this program. We shared together many moments and I am happy to have had you on my side to laugh and to support each other. I thank you because you helped me to grow, to become a better person, to enlarge my horizons and to think globally. Especially, a special thank goes to Roberto that even if he embodies the opposite side of me, is one of the most frank and honest friends (and cooks) I have. To Ezequiel, as proof that 2-years friendships can be strong as the life-long ones (even if you do not like Fiorentina steak). To my Mexicans friends, Sharon and Carmen, that stayed with me to laugh and to share good moments, always there to enjoy a piece of good cake (or ice-cream, or pasta, or pizza, or too-many-other-things) together. I would like to thank also Jose, always there to support

me and bear with me in Girona. It has been very nice meeting you all and I look forward to sharing more experiences.
My deepest appreciation goes to my family, my big family of 8 people plus the dogs. From distance, you have been always there to support me if I needed. I am so thankful for all the things you gave me and you taught me. Because if today I am glad of the person I have become, it is mainly because of what you made out of me.
Last but not least, I would like to thank Giovanni. Even if we are distant, he always stayed with me in my darkest moments, far but deeply present.

References

Banky, J.P., Kelly, J.W., English, D.R., Yeatman, J.M., Dowling, J.P., 2005. Incidence of new and changed nevi and melanomas detected using baseline images and dermoscopy in patients at high risk for melanoma. Archives of dermatology 141, 998–1006.

Barcelos, C.A.Z., Pires, V., 2009. An automatic based nonlinear diffusion equations scheme for skin lesion segmentation. Applied Mathematics and Computation 215, 251–261.

Benvenuto-Andrade, C., Dusza, S.W., Agero, A.L.C., Scope, A., Rajadhyaksha, M., Halpern, A.C., Marghoob, A.A., 2007. Differences between polarized light dermoscopy and immersion contact dermoscopy for the evaluation of skin lesions. Archives of dermatology 143, 329–338.

Chung, D.H., Sapiro, G., 2000. Segmenting skin lesions with partial-differential-equations-based image processing algorithms. IEEE transactions on Medical Imaging 19, 763–767.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine learning 20, 273–297.

Day, G.R., Barbour, R.H., 2000. Automated melanoma diagnosis: where are we at? Skin Research and Technology 6, 1–5.

Denton, W., Duller, A., Fish, P., 1995. Boundary detection for skin lesions: an edge focusing algorithm, in: Fifth International Conference on Image Processing and its Applications, IET. pp. 399–403.

Dhawan, A.P., Sim, A., 1992. Segmentation of images of skin lesions using color and texture information of surface pigmentation. Computerized Medical Imaging and Graphics 16, 163–177.

Fischer, P., Dosovitskiy, A., Brox, T., 2014. Descriptor matching with convolutional neural networks: a comparison to sift. arXiv preprint arXiv:1405.5769 .

Fornaciali, M., Carvalho, M., Bittencourt, F.V., Avila, S., Valle, E., 2016. Towards automated melanoma screening: Proper computer vision & reliable results. arXiv preprint arXiv:1604.04024 .

Friedman, R.J., Rigel, D.S., Kopf, A.W., 1985. Early detection of malignant melanoma: The role of physician examination and self-examination of the skin. CA: a cancer journal for clinicians 35, 130–151.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580–587.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2016. Region-based convolutional networks for accurate object detection and segmentation. IEEE transactions on pattern analysis and machine intelligence 38, 142–158.

Girshick, R.B., 2015. Fast R-CNN. CoRR abs/1504.08083.

Golston, J.E., Moss, R.H., Stoecker, W.V., 1990. Boundary detection in skin tumor images: An overall approach and a radial search algorithm. Pattern Recognition 23, 1235–1247.

Halpern, A.C., Marghoob, A.A., Bialoglow, T.W., Witmer, W., Slue, W., 2003. Standardized positioning of patients (poses) for whole body cutaneous photography. Journal of the American Academy of Dermatology 49, 593–598.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, in: Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE. pp. 2980–2988.

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, H., Bergstresser, P., 2007. A new hybrid technique for dermatological image registration, in: Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on, IEEE. pp. 1163–1167.

Kazhdan, M., Hoppe, H., 2013. Screened poisson surface reconstruction. ACM Transactions on Graphics (ToG) 32, 29.

Korotkov, K., Garcia, R., 2012. Computerized analysis of pigmented skin lesions: a review. Artificial intelligence in medicine 56, 69–90.

Korotkov, K., Quintana, J., Puig, S., Malvehy, J., Garcia, R., 2015. A new total body scanning system for automatic change detection in multiple pigmented skin lesions. IEEE transactions on medical imaging 34, 317–338.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection, in: CVPR, p. 4.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C., 2016. Ssd: Single shot multibox detector, in: European conference on computer vision, Springer. pp. 21–37.

Lowe, D.G., 1999. Object recognition from local scale-invariant features, in: Computer vision, 1999. The proceedings of the seventh IEEE international conference on, Ieee. pp. 1150–1157.

Matas, J., Chum, O., Urban, M., Pajdla, T., 2004. Robust wide-baseline stereo from maximally stable extremal regions. Image and vision computing 22, 761–767.

McConnell, R.K., 1986. Method of and apparatus for pattern recognition. US Patent 4,567,610.

Mcgregor, B., 1998. Automatic registration of images of pigmented skin lesions. Pattern Recognition 31, 805–817.

Menegola, A., Tavares, J., Fornaciali, M., Li, L.T., Avila, S., Valle, E., 2017. RECOD titans at ISIC challenge 2017. arXiv preprint arXiv:1703.04819 .

Mirzaalian, H., Hamarneh, G., Lee, T.K., 2009. A graph-based approach to skin mole matching incorporating template-normalized coordinates, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE. pp. 2152–2159.

Negin, B.P., Riedel, E., Oliveria, S.A., Berwick, M., Coit, D.G., Brady, M.S., 2003. Symptoms and signs of primary melanoma. Cancer 98, 344–348.

Perednia, D.A., White, R.G., 1992. Automatic registration of multiple skin lesions by use of point pattern matching. Computerized medical imaging and graphics 16, 205–216.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, pp. 91–99.

Roning, J., Riech, M., 1998. Registration of nevi in successive skin images for early detection of melanoma, in: Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on, IEEE. pp. 352–357.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Stewart, B., Wild, C.P., et al., 2017. World cancer report. Health .

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., et al., 2015. Going deeper with convolutions, Cvpr.

Tang, J., 2009. A multi-direction GVF snake for the segmentation of skin cancer images. Pattern Recognition 42, 1172–1179.

Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W., 2013. Selective search for object recognition. International journal of computer vision 104, 154–171.

Umbaugh, S.E., Moss, R.H., Stoecker, W.V., 1989. Automatic color segmentation of images with application to detection of variegated coloring in skin tumors. IEEE Engineering in Medicine and Biology Magazine 8, 43–50.

Umbaugh, S.E., Moss, R.H., Stoecker, W.V., 1992. An automatic color segmentation algorithm with application to identification of skin tumor borders. Computerized Medical Imaging and Graphics 16, 227–235.

Umbaugh, S.E., Moss, R.H., Stoecker, W.V., Hance, G.A., 1993. Automatic color segmentation algorithms-with application to skin tumor feature identification. IEEE Engineering in Medicine and Biology Magazine 12, 75–82.

Weinstock, M.A., 2006. Cutaneous melanoma: public health approach to early detection. Dermatologic therapy 19, 26–31.

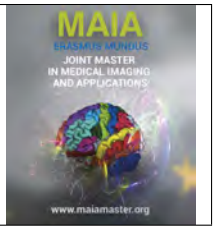
Wurm, E.M., Soyer, H.P., 2010. Scanning for melanoma. Australian Prescriber 33, 150–5.

Xu, L., Jackowski, M., Goshtasby, A., Roseman, D., Bines, S., Yu, C., Dhawan, A., Huntley, A., 1999. Segmentation of skin cancer images. Image and Vision Computing 17, 65–74.

Zhang, Z., Stoecker, W.V., Moss, R.H., 2000. Border detection on digitized skin tumor images. IEEE transactions on Medical Imaging 19, 1128–1143.

Medical Imaging and Applications

Master Thesis, June 2018



A Fully Automatic Framework for Myocardial Infarction Quantification in Late Gadolinium Enhancement MRI

Ezequiel de la Rosa, Désiré Sidibé, Alain Lalande

Le2i, Université Bourgogne-Franche Comté, Dijon, France

Abstract

Late gadolinium enhanced magnetic resonance imaging (LGE-MRI) is the gold standard and highest resolution technique for myocardial viability assessment. Although the technique accurately reflects the damaged tissue, there is no clinical standard for quantifying myocardial infarction (MI), demanding most algorithms to be expert dependent. Moreover, up to now the lack of public LGE-MRI datasets has hindered methods reproducibility and has restricted their objective comparison and validation. In this work we propose an end-to-end fully automatic framework for MI quantification in LGE-MRI. By means of a three-step deep learning based strategy, the framework provides automatic segmentation of the left ventricular myocardium, detection of diseased myocardial slices and subsequent infarction quantification. For its validation, reproducibility and further comparison against other methods, we developed a big multifield expert annotated LGE-MRI database. It accounts with healthy (n = 20) and myocardial infarcted (n = 80) scans and will potentially be opened in a public repository. In an exhaustive comparison against nine reference algorithms, the framework achieved state-of-the-art segmentation performances and showed to be the only method agreeing in volumetric scar quantification with the expert delineations. To our knowledge, this is the first fully automatic method for myocardial infarction quantification with high clinical transfer potential.

Keywords: Cardiac Magnetic Resonance, Late Gadolinium Enhanced, Scar Segmentation, Deep Learning

1. Introduction

Myocardial infarction (MI) is recognised by the World Health Organization as a severe global problem (Organization et al., 2011) and is the leading cause of death and main cause of morbidity in the US. It is produced by myocardial cell death as a consequence of prolonged ischemia, resulting in insufficient oxygen supply to some myocardium area (Rajiah et al., 2013). After the oxygen supply shortage, the affected myocardial regions can be non-viable or hibernating. Since in the hibernating regions the suspended myocardial activity is able to resume the contraction after revascularization (Wei et al., 2013a), viability assessment turns crucial for clinical and therapeutical decision making (Positano et al., 2005).

Cardiovascular magnetic resonance imaging (MRI) plays a key role in MI evaluation, since allows myocardial anatomy, left ventricular function, perfusion and viability assessment (Arai, 2008). The standard planes used in clinical practice are the short axis, horizontal long axis and vertical long axis of the left ventricle, from which the former one is the most common used for cardiac function evaluation (Petitjean and Dacher, 2011). Considering the long axis of the heart as the line extending from the apex to the center of the mitral valve, the short axis plane covers perpendicularly the long axis along the left ventricle. Most cardiac volumetric measurements on routine practice are conducted under the short-axis views (Ginat et al., 2011).

Late gadolinium enhancement (LGE) MRI is the cornerstone of myocardial tissue characterisation (Dastidar et al., 2015), representing the most accurate and highest resolution method for MI and non-ischemic cardiomyopathies diagnosis. It allows, as well, risk stratification and outcome prediction after revascularization processes or cardiac resynchronization therapy. Currently, LGE inversion recovery and phase sensitive inversion recovery are considered as the gold references for my-

ocardial viability assessment (Engblom et al., 2016). Imaging is conducted after 8-15 minutes of gadolinium injection which over-enhances infarcted myocardium by accumulation of the agent in the damaged tissue. In healthy tissue areas, given the fast gadolinium wash in and wash out no agent accumulation is presented and normal myocytes remain with hypointense signaling (Rajiah et al., 2013). By means of experimental studies, it was exhibited that the contrast distribution accurately reflects pathology of the myocardium (Kim et al., 1999). Infarcted tissue may also present hypointense regions as a consequence of the permanent microvascular obstruction (MVO, also call no reflow) phenomenon. MVO evidences the lack of reperfusion of some myocardial area even after the ending of the ischemic event, indicating severe ischemic disease and being associated with poor prognosis, adverse cardiac events and remodelling (Rajiah et al., 2013).

The main limitations of LGE-MRI for myocardial tissue assessment are not only due to technical parameters setting (such as slice thickness, inversion recovery, etc.) (Pattanayak and Bleumke, 2015) but mainly due to the lack of a clinical standard for scar tissue quantification (Engblom et al., 2016; Pattanayak and Bleumke, 2015). Thus, nowadays there is no reference method for abnormal tissue detection and segmentation, even though several techniques have been explored. It is worth to point out that excepting the STACOM 2012 challenge images (Karim et al., 2012), no open datasets can be found in this field, and hence most methods are only validated under private databases, hindering reproducibility and comparison against other techniques. The most frequently used techniques are the threshold-based ones, such as the full-width at half-maximum (FWHM) (Amado et al., 2004) and the n-standard deviations (from now n-SD) (Kim et al., 1999). Nevertheless, these methods provide poor agreement with expert delineations, inconsistent and high result variability and significant differences when compared it with ground truth (Spiewak et al., 2010; Zhang et al., 2016). Additionally, most of them are manual or semi-automatic requiring visual assessment and human interaction, turning the quantification process tedious, subjective and hardly reproducible. To our knowledge, up to now no fully automatized algorithm was proposed and current techniques require *i*) prior myocardium manual contouring or propagation from other MRI sequences (such as cine-MRI sequence) *ii*) visual identification of diseased myocardium before algorithm application (turning results valid for pathological slices only) and *iii*) scar tissue quantification by manual/semi-interactive methods.

In this work, the above-mentioned limitations are tackled by proposing an user interactive-free infarction quantification framework validated over a new cardiac database. The main contributions of this work are twofold. Firstly, it is presented a big multi-field LGE-MRI database for myocardial viability assessment. The free and open dataset accounts with fully annotated data including the left ventricular myocardium (LVM), blood pool, hyper-enhanced and no-reflow areas. To our knowledge, this is the first multifield dataset with expert delineated MVO areas. Secondly, herein it is proposed and validated an end-to-end, fully automatic framework for myocardial infarction quantification. The main novelties of this framework are: *i*) the automatic left ventricular myocardium segmentation on LGE images, *ii*) the discrimination of healthy and diseased myocardium slices, which extends the framework for working under healthy scenarios and *iii*) the development of a novel and robust automatic technique for scar tissue and MVO segmentation.

2. State of the art

2.1. Reference Datasets on Cardiac LGE-MRI

Unlike other medical image domains where well defined opened datasets allow algorithm performance quantification and direct results comparison, for the problem herein addressed such a reference proposal does not exist. Thus, methods reproducibility are mainly constrained to the private data used in the studies. Under the STACOM 2012 challenge a first attempt of a reference cardiac LGE-MRI database was presented accounting with pathological animal (15 cases, 3T magnetic field) and human (15 cases, 1.5T magnetic field) data (Karim et al., 2016, 2012). A big effort was conducted for performing various experts annotations, which was centered in hyper-enhanced tissue areas delineation but without providing microvascular-obstruction (MVO) ground truth. However, due to the limited database size as well as to the low quality of the acquired images (with many cases presenting overenhanced healthy areas), just few works extern to the challenge validated their algorithms with these images (Larroza et al., 2017). Besides, as far as the author knows until now there were no clinical studies conducted with different magnetic field MRI devices, a variable that may impact over algorithms performance.

2.2. Left Ventricular Myocardium Segmentation

Since scar tissue segmentation and quantification requires the initial delineation of the LVM (i.e., endocardium and epicardium contours have to be identified), in clinical practice is common to have expert manually drawn boundaries. According to Petitjean and Dacher (2011), this task requires around 20 minutes for a clinician in Cine-MRI (half time on LGE-MRI), turning the process long and tedious. The left ventricular myocardium segmentation in cardiac MRI presents well known difficulties (Bernard et al., 2018; Petitjean and Dacher, 2011): *i*) Poor contrast between the epicardium and the surrounding structures (and high-contrast between the blood-pool and the endocardium), *ii*) Gray-level inhomogeneities in the left ventricular cavity due

Table 1: Summary of published methods for scar and left-ventricle myocardium segmentation.

					Scar Segmentation		LVM Segmentation	
Reference	**DB**	**n**	**Seq**	**MF**	**Algorithm**	**Inter**	**Algorithm**	**Inter**
Kim et al. (1999)	A	26	IR	1.5T	2-SD	Semi	-	-
O'Donnell et al. (2003)	H	14	IR	-	SVM	Auto	Expert Drawing	Man
Dikici et al. (2004)	H	45	IR	-	Intensity Features & SVM	Auto	Cine Propagation	Semi
Amado et al. (2004)	A	30	IR	1.5T	n-SD, FWHM	Semi	Expert Drawing	Man
Kolipaka et al. (2005)	H	23	IR	1.5T	n-SD	Semi	Expert Drawing	Man
Positano et al. (2005)	H	15	IR	1.5T	Fuzzy Clustering	Auto	Drawing & Active Contours	Semi
Hsu et al. (2006)	A	11	PSIR	1.5T	EM, CCA & FWHM	Auto	Expert Drawing	Man
Schmidt et al. (2007)	H	47	IR	1.5T	n-SD + FWHM	Semi	Expert Drawing	Man
Ciofolo et al. (2008)	H	27	-	-	-	-	Template Deformation & Cine Propagation	Semi/ Auto
Hennemuth et al. (2008)	H	21	IR	-	EM + Watershed	Auto	Live-Wire	Semi
Detsky et al. (2009)	H	15	IR	1.5T	Fuzzy clustering	Auto	Expert Drawing	Man
Tao et al. (2010)	H	20	IR	1.5T	Otsu, CCA & Region Growing	Auto	Expert Drawing	Man
Andreu et al. (2011)	H	12	IR	3T	50, 60, 70% FWHM	Semi	Expert Drawing	Man
Flett et al. (2011)	H	60	IR	1.5T	n-SD, FWHM	Semi/ Auto	Expert Drawing	Man
Valindria et al. (2011)	H	20	PSIR	3T	EM, FWHM & Feature Analysis	Semi	Expert Drawing	Man
Lu et al. (2012)	H	10	IR	1.5T	Graph-cut, EM & FWHM	Auto	Cine Propagation & Manual Correction	Semi
Wei et al. (2013b)	H	21	IR	1.5T	-	-	Cine Propagation, Mesh Deformation & Intensity Modelling	Semi
Wei et al. (2013a)	H	20	IR	1.5T	3D Graph-Cut	Auto	Cine Propagation, Registration & Intensity Modelling	Man/ Semi
Kotu et al. (2013)	H	44	-	1.5T	Probability Maps	Semi/ Auto	Expert Drawing	Man
Pop et al. (2013)	A	9	IR	1.5T	EM	Auto	-	-
Rajchl et al. (2014)	H	50*	IR	3T	Potts Models	Semi	-	-
Engblom et al. (2016)	A/ H	38/ 124	IR/ PSIR	1.5T	EM, Weighted Intensity & A-priori Information	Auto	Expert Drawing	Man
Ukwatta et al. (2016)	H	61	IR	1.5T	Min-Cut Optimization	Auto	Expert Drawing	Man
Liu et al. (2017)	H	22	-	-	-	-	GMM & Level Set	Auto
Kruk et al. (2017)	H	7	PSIR	3T	Watershed & Shape Priors	Semi	Expert Drawing	Man
Kurzendorfer et al. (2017a)	H	100	IR	1.5T	-	-	Hough transforms, Active Contours & Random Forests	Auto
Kurzendorfer et al. (2017b)	H	30	IR	3T	-	-	Registration, Polar-Space Refinement & Marching Cubes	Auto
Xu et al. (2017)	H	114	-	3T	Deep Learning	Auto	-	-

Note: DB: Database; Seq: MRI Sequence; MF: Magnetic Field; LVM: Left Ventricular Myocardium; Inter: Interaction; A: Animal; H: Human; IR: Inversion Recovery; PSIR: Phase-sensitive IR; SD: Standard-deviation; FWHM: Full-width at Half-maximum; EM: Expectation-maximization; CCA: Connected Component Analysis; GMM: Gaussian Mixture Model; Man: Manual; Semi: Semi-automatic; Auto: Automatic.

to the blood flow, *iii*) Presence of papillary muscles and trabeculations in the chamber, with similar gray-levels as the myocardium, *iv*) partial volume effect given by MRI limited resolution, *v*) Motion artifacts given by respiration and heart dynamics, *vi*) shape and gray-level heterogeneity over the heart structures by patient and pathology and *vii*) Presence of banding noise. It is consequently highly desired the devising of automatic segmentation methods that can perform the left-ventricle delineation dealing with the above presented difficulties. In this sense, the task was widely addressed in Cine MRI sequences, with over seventy proposals listed by Petitjean and Dacher (2011) and several newer deep-learning strategies (Avendi et al., 2016; Bernard et al., 2018; Curiale et al., 2017; Oktay et al., 2018). However, over LGE-MRI sequences the task was much less explored and remains being an open issue. It is worth to point out some additional constraints presented in contrast-enhanced MRI that does not allow using the same strategies developed for kinetic sequences. Firstly, the images present high myocardial gray-level heterogeneity in infarcted images, while intensity homogeneity is preserved in non-diseased ones. Secondly, images present noisier and less sharpen organ boundaries when compared against Cine MRI. For this reason, the signal contrast between the healthy myocardium and the lungs is poor. Thirdly, scarred myocardial tissue presents similar intensity levels as the blood-pool (Tao et al., 2010) and both structures are generally contiguous (infarction propagates from the endocardium to the epicardium). Given these characteristics of cardiac LGE-MRI, the myocardium delineation turns extremely challenging.

The few algorithms proposed until now for segmenting the myocardium in LGE-MRI are summarized on Table 1. As can be appreciated, most approaches are semi-automatic and/or require a-priori Cine MRI segmentations. Despite having strong limitations, propagation of the LVM boundaries from Cine MRI was widely explored (Ciofolo et al., 2008; Dikici et al., 2004; Lu et al., 2012; Wei et al., 2013a,b). Hennemuth et al. (2008) proposed using the live-wire-algorithm for LVM segmentation and Liu et al. (2017) used multi-component Gaussian mixture model and coupled level sets. Kurzendorfer et al. (2017a) used a random forest and dynamic programming based algorithm and later extended the work to 3D LGE-MRI by using a multiple-step registration, polar-space refinement and marching cubes algorithm (Kurzendorfer et al., 2017b). Nonetheless, while semiautomatic techniques require expert-interaction or a-priori kinetic sequences information, automatic algorithms are still limited in terms of performance, speed and validation. Thus, until now there is no consensus of a reference method for segmenting the left ventricular myocardium in LGE-MRI.

2.3. Myocardial Damage Detection

Quantification of myocardial scar has been widely addressed in many works where most algorithms were mainly validated under pathological datasets (Table 1). Before their application for scar quantification, myocardial abnormality identification by visual inspection is a compulsory step, as in n-SD (Kim et al., 1999) and FHWM methods (Amado et al., 2004). By means of this approach, the lesions search is guaranteed in abnormal myocardiums only, reducing false positives detection in healthy slices. Despite the fact that these methods are able to control better the false positive rate, one of the drawbacks is the expert interaction required. Under this scenario the development of an automatic adaptive framework that could deal with healthy patients as well turns highly desirable.

Devising such a method based on intensity myocardial profiles could be conducted by characterizing healthy and abnormal myocardium histograms. In previous works, healthy and scarred myocardial tissue distributions have been well described (Hennemuth et al., 2008; Tao et al., 2010; Wei et al., 2013a). While a Rayleigh (or a Rician) distribution might appropriately model the normal tissue, hyper-enhanced infarcted areas are suitable modeled by a Gaussian one. Thus, the whole myocardium histogram consists on the resulting distribution obtained from the overlapping of healthy and abnormal tissues. For the sake of simplicity, assumption of both distributions as Gaussian models has been extensively conducted (Carminati et al., 2016; Engblom et al., 2016; Pop et al., 2013; Valindria et al., 2011). Hennemuth et al. (2008) proposed the use of information criteria (Akaike and Bayesian ones) for histogram characterization. By assessing the best model fitness of an histogram could be distinguished weather the myocardium is normal (best fitness achieved with only one distribution) or abnormal (best fitness with two, overlapped distributions). However, the main limitation of

Table 2: MRI acquisition parameters summary.

Sequence	MF	TE	FA	Matrix (min/max)	Spatial Resolution (min/max)	ST	SG
PSIR	1.5T	1.37 ms	25°	240 x 138 / 256 x 224	1.87 x 1.87 / 1.25 x 1.25 mm^2	8 mm	2 mm
PSIR	3T	1.53 ms	20°	256 x 168 / 256 x 256	1.91 x 1.91 / 1.36 x 1.36 mm^2	8 mm	2 mm

Note: PSIR: Phase Sensitive Inversion Recovery; MF: Magnetic Field; TE: Echo Time; FA: Flip Angle; ST: Slice Thickness; SG: Slice Gap.

this approach regards expectation-maximization algorithm convergence. Due to the considerable distributions overlap, the algorithm sometimes converge to a unique component model. Evenmore, in small myocardial lesions, the scarred tissue distribution is obscured by the healthy one, turning the method inaccurate.
The problem could be better addressed by using not only intensity features. In this sense, recently Zreik et al. (2018) proposed a method for coronary artery stenosis detection by myocardial characterization in CT scans. The method extracts local myocardial features using convolutional autoencoders and through clustering and support-vector-machines (SVM) classifies healthy and abnormal slices. Despite the good performance achieved, implementation of this approach turns difficult in clinical practice due to the extensively long computational time required. Even more, since the algorithm was developed for CT scans there is no warranty that could work on LGE-MRI.

2.4. Myocardial Infarction Segmentation

For a better understanding of the existing algorithms for MI segmentation, we summarized the most relevant works on Table 1. As can be appreciated, intensity based segmentation algorithms have been widely investigated and validated in clinical practice. In these techniques, the histogram thresholding is conducted in a semiautomatic approach by a myocardial region delineation done by an expert (like in n-SD, Kim et al. (1999); Kolipaka et al. (2005), and FWHM methods, Amado et al. (2004)) or in an automatic way by an optimized cutoff value selection (like in expectation-maximization, Pop et al. (2013) or Otsu (1979) approaches). Given that these methods can not deal with the overlapping tissue distribution areas, several studies extended or combined these methods by using more sophisticated tools. Common works recombined the thresholding techniques (Andreu et al., 2011; Flett et al., 2011; Schmidt et al., 2007) or used intensity features with connected component analysis (Hsu et al., 2006; Tao et al., 2010; Valindria et al., 2011), clustering (Detsky et al., 2009; Positano et al., 2005) or SVM (Dikici et al., 2004; O'Donnell et al., 2003). Graph-cuts (Lu et al., 2012; Wei et al., 2013a) and watershed algorithms (Hennemuth et al., 2008; Kruk et al., 2017) have received researchers attention as well.
Despite the vast techniques exploration, up to now there is no reference method for scar quantification (Engblom et al., 2016) and just few of these techniques are used in clinical practice. The considered state-of-the-art (SOA) ones comprise the n-SD and FWHM, even when its variability, reproducibility and lack of expert agreement was highly discussed (Spiewak et al., 2010; Zhang et al., 2016). For these reasons, the development of a robust technique able to accurately reproduce the experts delineations turns highly valuable. With these aims, Xu et al. (2017) attempted the problem by proposing a solution under the deep learning paradigm. The poor performance assessment by the only use of classification metrics as well as the lack of appropriate validation makes the method hardly usable in clinical practice.

3. Material and methods

3.1. Data Acquisition

A cohort of healthy and myocardial infarcted patients that attended the imaging center of the University Hospital of Dijon (Dijon, France) between 2015 and 2017 were included in the study. The institutional review board approved the study development. A total amount of 100 randomly chosen late-gadolinium enhanced MRI cases (20 healthy, 80 presenting infarction) were considered for developing the dataset. Gadolinium contrast solution (Dotarem, Guerbet, France) was administered to the patients between 8 to 10 minutes before conducting the study. Myocardial infarction was assessed and confirmed in all cases by LGE-MRI. The 35% of infarcted cases ($n = 28$) presented micro-vascular obstruction areas detected as hypo-enhanced zones surrounded by enhanced necrotic tissue. For all patients, a short-axis stack of cardiac images covering the whole left ventricle were acquired using one of the two clinical MRI devices with magnetic fields of 1.5T and 3T (Siemens Medical Solutions, Erlangen, Germany). All volumes account with between 6 to 9 images, with a slice thickness of 8 mms and a distance between two slices of 10 mms. Voxel resolution varied according to the patient between 1.25 x 1.25 mms^2 to 1.91 x 1.91 mms^2. For each patient, a phase sensitive inversion recovery sequence was used for acquiring the images. All datasets were stored using the digital imaging and communications in medicine (DICOM) format and anonymized for research purposes. A summary of the acquisition sequence parameters used is shown on table 2.
The dataset ground truths were delineated in each slice by an expert of the institution (AL) with more than 10 years of expertise in the field. The endocardium and epicardium boundaries were contoured, defining the left ventricular myocardium and blood-pool regions. Papillary muscles were included in the cardiac cavity, as recommended (Lalande et al., 2015). Afterwards, in pathological cases the scar tissue was annotated taking separate contours for enhanced and microvascular obstructed areas. The full database accounts with 751 images from which 474 presented diseased myocardium (111 including microvascular obstructions) and 277 were healthy. For assessing intra/inter-observer annotations variability, a random subset (50%, $n = 40$) of pathological cases were re-contoured by the same expert as well as by a second observer (Dr. Leclerq, a cardiologist with 5 years of experience in the field).
I fully developed the dataset as part of this master thesis project. The activities performed include old data

retrieving, cleaning, MI and MVO cases selection, repeated slices or missing information cases removal, slices re-ordering (base to apex), delineation inaccuracies detection, dataset statistics computation, scans format convertion (DICOM to NIfTI) and storing.

3.2. Proposed Framework

In this work a fully automatic framework for detection and quantification of myocardial infarction from short-axis cardiac LGE-MRI is proposed. The algorithm receives as input raw MRI scans and in a 3-step algorithm approach provides the left-ventricle myocardium segmentation, per-slice detection of diseased heart and quantification of damaged myocardial regions. The outputs of the framework are the the segmented myocardium, the detected scarred areas and their corresponding clinical markers of medical importance. Firstly, myocardium segmentation is conducted under a deep-learning strategy which defines the myocardial region of interest. Secondly, by mimicking the experts clinical working pipeline healthy and pathological scans are discriminated before conducting segmentation. Developing such a method benefits the infarction segmentation performance, since avoids potential lesion over-estimation by false-positive inclusions in healthy images. Thirdly, the scarred tissue is quantified by an initial fast coarse segmentation followed by a voxel reclassification refinement strategy. The whole framework was implemented under Python and Matlab® R20017b environments.

3.2.1. Data Pre-processing

Collected MRI scans present differences among them mainly in *i*) voxel size and *ii*) intensity values. While the former differences come from the setting of diverse scanning parameters (which are decided depending on the examination by the involved device operator), the latter differences may come from the use of different magnetic field devices (which account with diverse signal to noise levels) as well as by the inherent biological and anatomical patients variability. For homogenizing the scans and addressing data variability all volumes were pre-processed by following three steps. Firstly, high-frequency noise was removed from all slices by using a spatial adaptive non-local means filter with automatic noise level estimation (Manjón et al., 2010). The chosen algorithm allows to tackle not only the intra-patient noise level differences in the scan, but also the inter-patient one observed by the use of different MRI magnetic field devices. Secondly, volumes were re-sliced to reach an homogenous voxel size of 1.25 x 1.25 x 8 mm^3 (minimum voxel space found among patients) by means of linear interpolation. Thirdly, for reducing inter-patient intensity variability, all images were normalized within the interval [0-1].

3.2.2. Left Ventricular Myocardium Segmentation

Segmentation of the left ventricle myocardial boundaries was conducted by means of the 2D U-Net convolutional-neural-network (CNN) winner of the ISBI 2012 challenge (Ronneberger et al., 2015). This architecture showed outstanding performances for segmenting the myocardium on kinetic MRI under the ACDC challenge (Bernard et al., 2018), where the two best ranked groups used U-Net based approaches (Baumgartner et al., 2017; Isensee et al., 2017). The preference of a 2D architecture instead of the 3D one relies on the poor resolution along the Z-axis and on the variability of the diafragm position during two consecutive breathholds. Thus, 3D architectures could help in myocardial segmentation by using strategies for dealing with the motion constrains. Nonetheless, compensation of breathholding variabilities requires developing more complex models, which are out of this work scope. Our implementation differed from the original work of Ronneberger et al. (2015) since padded convolutions were used for keeping dimension consistence across the concatenation levels. Moreover, our proposal differs from the works of Isensee et al. (2017) and Baumgartner et al. (2017) since we followed a patch based segmentation strategy instead of working with the entire images.

Training phase. Patches of size 64 x 64 were used for segmenting the left ventricular myocardium, size choosen on a preliminary patch dimension performance exploration. Given that the myocardial anatomy represents a small part of the whole MRI slices, dealing

Table 3: Summary of the CNN architectures used in the framework and their corresponding parameters.

Goal	Net	Patch	LF	Optim	LR	M	MB	L2	E	ES
LVM Segmentation	U-Net	64x64	CE	Adadelta	1	-	32	-	50	✓
Disease Detection	VGG19	89x89	CE	SGDM	$1x10^{-4}$	0.9	16	$1x10^{-4}$	20	×
Scar Segmentation	Zreik et al. (2018)*	49x49	CE	SGDM	$1x10^{-2}$	0.75	256	$1x10^{-4}$	50	×

Note: LVM: Left Ventricular Myocardium; Net: Network Architecture; LF: Loss Function; Optim: Optimizer; SGDM: Stochastic Gradient Descent with Momentum; LR: Learning Rate; M: Momentum; MB: Minibatch Size. L2: L2 Regularizer; E: Epochs; ES: Early Stopping. * The elemental branch of the network was used instead of the whole architecture.

with class imbalance turns crucial for avoiding a background biased segmentation. For tackling this problem, the training was conducted under a class balanced patch extraction strategy were 50 % of the patches were randomly centered in background areas and the remaining 50 % were centered in myocardial areas. A high information patch overlap was considered during the extraction step, since helps in the network learning process (Bernal et al., 2018). Moreover, for avoiding information redundancy among the overlapped patches and with the aim of increasing the CNN learning capability we added random offsets to the central patch voxels (Guerrero et al., 2018). This strategy allows the possibility of finding myocardial voxels elsewhere the patch, and not only in the patch center. Moreover, for overfitting avoidance we conducted data augmentation for increasing the training set by considering random geometric image transformations (rotation, shearing, flipping and scaling). On Table 3 the used network training parameters are summarized.

Segmentation phase. For conducting the voxels label prediction we performed overlapping patch extractions along the whole image. Afterwards, each patch was independently pass by the trained network and the predicted labels were obtained. The final segmentation was conducted, for each voxel, following a maximum-a-posteriori probability approach were the label probabilities from the different patches were equally-weighted combined.

3.2.3. *Myocardial Abnormality Detection*

Once the contours that enclose the myocardium are found, we are interested in knowing the condition of the myocardial tissue in each slice. For it, a dichotomous classifier is built for discriminating healthy and pathological images. By using the epicardial mask and by assuming that the myocardial heart shape resembles a ring, the epicardium centroid it is estimated. Afterwards, cropped images (size 89x89, 3-channel replicated) masked within the myocardium and centroid-centered are used as inputs of the classifier.

Classification phase. For achieving the classification task a three-step approach is conducted: *i*) Fine-tuned VGG19 (Simonyan and Zisserman, 2014) models are used for extracting informative features characterizing the myocardial images, *ii*) Extracted features followed a principal-component-analysis (PCA) dimensionality-reduction and *iii*) images are finally classify as healthy or infarcted by using support-vector-machines. In step *i*), the images to classify are passed through the fine-tuned neural network and the 1000 features from the last fully connected layer (FCL) are extracted. Afterwards, in step *ii*) the features are projected into the principal-components space after mean-feature vector subtraction, obtaining k new features ($k \ll 1000$) which preserve 95% of the data variance. The classification stage (step *iii*) is conducted on a linear kernel trained support-vector-machine by using the k principal component features.

Training phase. The ImageNet pretrained VGG19 (Simonyan and Zisserman, 2014) model was chosen over other network architectures (such as VGG16, Resnet50, Resnet101, and GoogleNet) based on an exploratory performance analysis. In previous works, the model shows suitability and good adaptability for working in the medical domain (Antropova et al., 2017; Jia et al., 2018). Since the main aim of this framework block is the devising of a robust image classifier, only

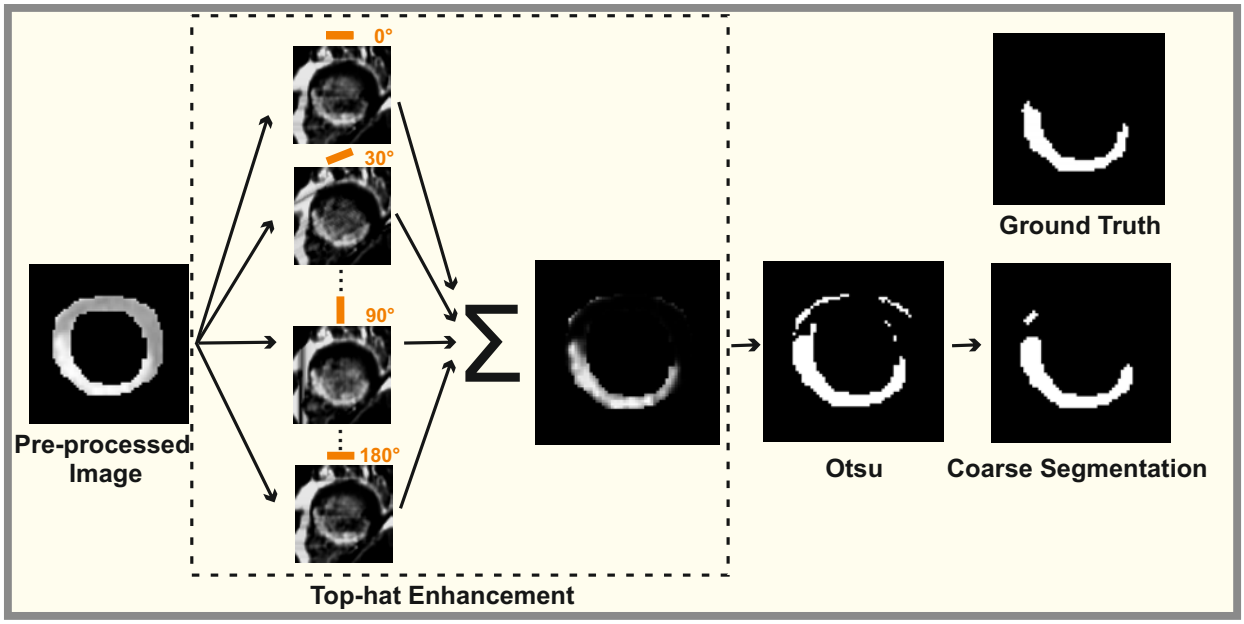


Figure 1: Coarse segmentation workflow.

experiments with the achieved outperforming network are shown. The model was fine-tuned using MR images by preserving all layers and its corresponding weights with exception of the three ending FCL's, whose neurons weights where re-learnt. Besides, after the last 1000-neuron FCL layer an extra 2-neuron FCL with a softmax layer were added for conducting the classification. The network training parameters are summarized on Table 3. For the replaced FCL, the learning rate was 30 times higher the value shown on the table. Considering the dataset size limitations and with the aim of avoiding overfitting we *i*) performed data augmentation as explained in Section 3.2.2, *ii*) shuffled the training set in every epoch and *iii*) applied a random dropout [Srivastava et al., 2014] of 50% after each fully connected layer. Besides, for avoiding the classifier to produce biased class results, data imbalance was addressed by randomly sub-sampling the majority class until reaching the minority class size.

The PCA step was conducted as in (Sidibe et al., 2017). Features extracted from the 1000 FCL over the training set were used for building a matrix $\mathbf{X} = [b_1\ b_2\ \cdots\ b_m]$, where m is the amount of training samples and $b_i \in \mathbb{R}^{1000}$. The mean feature vector was computed as $\bar{b} = \frac{1}{M}\sum_{j=1}^{M} b_j$ and was subtracted from each column of $\mathbf{X}$ for centering the data. Afterwards, the covariance matrix was computed as $\mathbf{C} = \frac{1}{M}\mathbf{X}\mathbf{X}^{\mathbf{T}}$ and its eigendecomposition was conducted such that $\mathbf{C} = U \wedge U^T$. Thus, the principal components are the eigenvectors (columns) of $\mathbf{C}$ and the eigenvalues λ_i ($i = 1, \cdots, 1000$) give the axis importance. For conducting data-dimensionality-reduction we kept the k principal components associated with the k largest eigenvalues, such that 95% of the data variance is preserved (condition satisfied when $\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{1000}\lambda_i} > 0.95$). Afterwards, with the dimensionality-reduced training set $\mathbf{X'} = [b'_1\ b'_2\ \cdots\ b'_m]$ (where $b'_i \in \mathbb{R}^k$) we fit a SVM with linear kernel for classifying normal and infarcted myocardium images.

Model Validation. One-hundred random dataset splits were conducted in a class balanced 80-10-10% (training-validation-test) approach. For each training-validation set, fine tuning of VGG19, principal components decomposition and SVM fitting were conducted. Afterwards, over the test-set the label prediction was performed. Obtained classifiers were characterized and evaluated by means of a receiver-operating-characteristic (ROC) curve analysis. Besides, to further validate whether the discriminant SVM rule could be randomly achieved, a one-hundred permutation analysis (Ernst et al., 2004) over *Healthy vs Diseased* cases was performed. Obtained ROC area under the curve (AUC) values were used as a global performance metric for comparing permuted and un-permuted classifier results.

3.2.4. Myocardial Scar Segmentation

Once abnormal slices were detected in the volume, the lesion segmentation was conducted in a two-steps approach by firstly providing a fast coarse segmentation and secondly by refining it using a voxel-reclassification strategy. Images were enhanced by means of gamma correction $I_{out} = I_{in}^{\gamma}$ ($\gamma > 1$) and normalized for having a distribution within the epicardium inner region spread in the interval [0-1]. Since the contrast agent tissue concentration changes within time and the intensities become brighter from the mitral valve to the apex causing inter-slice variability (Wei et al., 2013a), all images were normalized by using the left ventricular myocardium and blood-pool regions information. By means of this approach the myocardial intensity homogeneity was guaranteed for different slices and patients.

Coarse Segmentation. Myocardial infarction was initially segmented by a non-parametric intensity based approach applied after over-enhancing potential damaged regions, as similarly conducted on Ram et al. (2012), BahadarKhan et al. (2016) and Savelli et al. (2017) in other fields. The enhancement was conducted by using a sum of non-linear top-hat transforms, which increases the contrast between dark and bright image areas (healthy and damaged regions respectively). The top-hat transform is defined as:

$$g = I - (I \circ B) \tag{1}$$

where g is the transformed input image I and $\circ$ represents the gray-scale opening operator using a structuring element. The transform was applied in each slice using a 2D bar rotational structuring element with increasing variations of 30° and a constant length of 34 pixels. The top-hat enhancement was conducted for reducing the overlapping areas of the healthy and scar tissue intensity distributions due to partial volume effect, helping the tissues discrimination by the automatic thresholding Otsu algorithm (Otsu, 1979). Structuring element shape, length and rotations degree were empirically selected by maximizing the segmentation performance over the training set. Subsequently, a morphological opening (disk as structuring element, radius 1 pixel) was applied for removing small misclassified voxel clusters. The coarse segmentation workflow is shown on Fig. 1.

Refined Segmentation: I)Voxel reclassification phase. After achieving an initial segmentation of the potential damaged areas, a voxel-level segmentation refinement was followed by using an ensemble of from-scratch trained CNNs. Although in the initial segmentation most lesions are detected within their core or more evident damaged areas, the method might provide misclassifications due to the overlapping healthy and infarcted intensity distributions. Thus, we tackle false positives (respectively negatives) removal (resp. inclusion) by a voxel reclassification approach using image patches

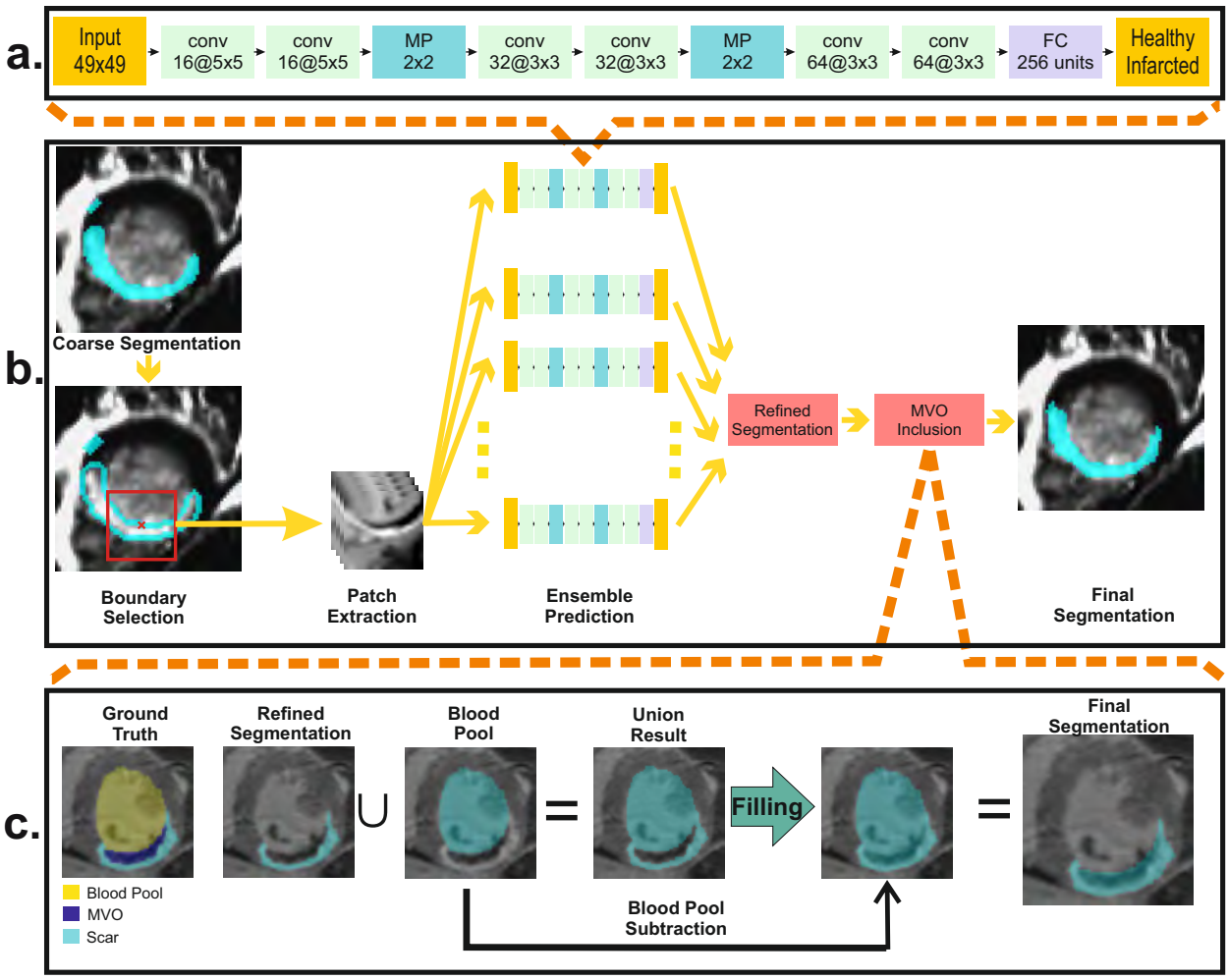


Figure 2: Segmentation refinement block. a) CNN architecture used in the ensemble. b) Refinement segmentation workflow. c) Microvascular-obstruction inclusion workflow. conv: Convolutional Layer; MP: Max-Pooling Layer; FC: Fully Connected Layer.

centered at the voxels of interest. Even if an infarction might affect the whole myocardium in a slice, the amount of voxels to reclassify represent a small amount of data compared with the whole MRI scan. Under this assumption, we reclassify voxels falling in a boundary region surrounding the pre-segmented mask after morphologically dilating it and eroding it (disk as structuring element with radius of 2 pixels) with the aim of including in the analysis potential misclassified voxels surrounding the initial segmentation. The degree of the dilation was chosen by assuring at least a mean sensitivity of 95% over the dilated mask on the training set. Voxel label prediction was achieved afterwards by majority voting after passing a patch (centered on the voxel of interest) over a seven-component CNN ensemble. The whole refinement segmentation workflow can be appreciated on Fig. 2.b).

Refined Segmentation: II) CNN's training phase. The CNN architecture of Fig. 2.a) was used for voxel classification, which consists on a modified single-branch version of the architecture proposed for left ventricular myocardium segmentation in Zreik et al. (2018). Unlike the original implementation, we used rectified linear units as activation functions (Krizhevsky et al., 2012). After each convolutional layer, batch-normalization (Ioffe and Szegedy, 2015) was used with the aim of rendering the training process faster and less sensitive to learning rates (Zreik et al., 2018). For building the classifier, the network was trained from scratch by using patches taken from the training set in a 50%-50% class balanced way. Ground-truth masks were dilated and eroded by using a disk of radius 5. Then, healthy-class patches were taken from the mask obtained after subtracting to the dilated mask the original ground-truth. Likewise, infarcted-class patches were extracted from the mask obtained after subtracting to the original infarction mask the eroded mask. In cases were the lesions were small (and hence the erosion operation degraded the whole mask) patches from the entire mask were taken. The reason for preferring boundary-close voxels instead of central ones relies on the difficulty for their detection, since partial volume effect and the so

call gray-zone areas (Valindria et al., 2011) turns the tissue separation difficult. Voxel-centered patches were extracted with a stride of 3, providing high information overlap.

A summary of the parameters used during the training phase is shown on Table 3. Hyper-parameters selection was conducted by an exhaustive sequential manual search that lead to segmentation performance maximization. A patch size of 49x49 was chosen, which provided enough contextual information for learning low as well as high level features. In all cases, patches were zero-centered by subtracting the mean image of the training set. Overfitting avoidance and data balancing were conducted as before described (Section 3.2.3). The network training process converged after 50 epochs.

The classifiers ensemble was built by training CNNs in a 7-fold cross-validation strategy (over the considered training set) where networks were trained in the same fashion. Thus, 7 CNN's were obtained per each training set for conducting voxel reclassification by majority voting over the test-set. Validation of the method was performed by 5-fold cross-validation (80-20% of patients as training-test sets respectively in each fold).

Table 4: Dataset.

Characteristic	All	Patients Healthy	Diseased HE	Diseased MVO
n	100	20	52	28
Age	59.5 (12.8)	53.4 (15.2)	61.1 (11.0)	62.4 (13.7)
Men	62	14	38	10
Women	20	5	12	2
$Scar_{vol}$ (mm^3)	25.6 (19.3)	-	22.5 (19.2)	31.5 (18.8)
LVM_{vol} (mm^3)	225.1 (87.4)	232.2 (57.3.7)	221.9 (73.0)	207.8 (52.9)
IM (%)	18.5 (12.6)	-	15.4 (11.6)	24.3 (12.7)
Subset (50%)				
$Scar_{vol}$ (mm^3)	40.7 (32.1)	-	-	-
LVM_{vol} (mm^3)	211.2 (76.5)	-	-	-
IM (%)	18.9 (13.8)	-	-	-

Note: Mean (standard deviation). IM: Infarcted Myocardium. LVM: Left Ventricular Myocardium; HE: Hyper-enhanced Areas; MVO: Microvascular Obstructed Areas. *Age* and *Sex* fields were field with available information only.

Refined Segmentation: III) MVO Inclusion. The infarction segmentation algorithm developed only accounts for hyper-enhanced regions detection. For including micro-vascular obstruction areas, we took advantage of pathological anatomy a-priori information since MVO is represented as hypointense regions neighbouring the hyperintense areas (Durante and Camici, 2015). Besides, infarction is always propagated from the endocardial cavity towards the epicardial one (Rajiah et al., 2013), assuring connectedness of the enhanced scar tissue volume with the blood-pool area. Mainly, MVO is found in the images as a dark cluster of voxels *i*) confined by the endocardial and enhanced areas or *ii*) fully enclosed in the enhanced region. For MVO inclusion, we computed the union of the endocardial and hyperintense infarction masks for finding all voxels clusters fulfilling *i*) and/or *ii*). Afterwards, holes were filled and the endocardial mask was finally removed for having a unique infarction segmentation mask including dark and bright pixel areas. The MVO inclusion algorithm in illustrated on Fig. 2.c).

Comparison against SOA methods. In order to evaluate the proposed infarction segmentation algorithm performance, results were compared against nine standard algorithms (including SOA ones) widely used in clinical practice: the n-SD ($n = 1, 2, \cdots, 6$) (Kim et al., 1999), Otsu (Otsu, 1979), FWHM (Amado et al., 2004) and Gaussian Mixture Models (with threshold at 2-SD above the mean healthy intensity) (Pop et al., 2013). All SOA algorithms were from-scratch implemented in this work following the original paper implementations.

3.2.5. Performance & Statistics

Statistical analysis were conducted over the different estimated metrics by first inspecting data behaviour and then applying one of the following tests: (un)-paired t-Student test, non-parametric Wilcoxon and Mann-Whitney U-tests (paired and unpaired respectively). For t-Student test conductions, normality was checked by using the Shapiro-Wilk test while homoscedasticity was verified by data distribution inspection. When these requirements were not fulfilled, non-parametric tests were preferred. Two-tailed tests with a 0.05 significance level were performed in all cases.

Performance & Statistics: I) Dataset. The dataset summary is reported by its mean and standard deviation variable values. For assessing the observers delineation agreement, intra/inter-observer variabilities were computed in terms of left ventricle myocardial volume (cm^3), scar volume (cm^3) and percentage of scarred myocardium ($\%\frac{Vol_{scar}}{Vol_{Myocardium}}$) by using Spearman correlation and Bland-Altman analysis (Bland and Altman, 1986).

Performance & Statistics: II) Classification. Image classification was assessed by the methods sensitivity, specificity and accuracy. Besides, characterization of the built classifier was evaluated by the area under the

curve on the ROC analysis. Mean and standard deviation of AUC values were reported. For assessing the model robustness in the ROC permutation analysis, the $p - value$ was computed as follows:

$$p = \sum_{i=1}^{N} \frac{I(AUC_i^p, AUC_i^{np})}{N} \quad (2)$$

where $N = 100$ is the amount of data splits (and permutations) conducted, AUC_i^p and AUC_i^{np} are the obtained AUC values for the permuted and un-permuted $i - th$ dataset split respectively and the indicator function I follows the below defined rule:

$$I(AUC_i^p, AUC_i^{np}) = \begin{cases} 1 & if\ AUC_i^p \geq AUC_i^{np} \\ 0 & if\ AUC_i^p < AUC_i^{np} \end{cases} \quad (3)$$

Performance & Statistics: III) Segmentation. For the different segmentation goals and for all the algorithms compared the performances were assessed by Dice similarity indexes ($Dice = 2\frac{|X \cap Y|}{|X|+|Y|}$, being X the obtained volume and Y the ground truth) and 3D Haussdorf distances. The total myocardial volume, scarred myocardial volume and percentage of infarcted myocardium were quantified for assessing the performance of clinical markers estimation. Furthermore, in order to assess over/under-segmentation of the different algorithms, the relative volume differences (RVD) were computed as $RVD = \frac{|X|}{|Y|} - 1$. In addition, for all segmentations and methods the estimated volumes were compared with the expert annotated ones by using Spearman correlation coefficient and Bland-Altman plots (mean and standard deviation of bias are provided).

4. Results

4.1. Dataset

A summary with the whole dataset retrieved clinical information as well as with the estimated markers of clinical interest in shown on Table 4 (superior part). Likewise, results for the 50% random chosen cases used for intra/inter-observer variability are summarized on the inferior part of the table.

When the re-delineation of the data subset was conducted by the main observer, the corresponding results of Table 5 (superior part) were obtained. On the other hand, in the inferior part of the table the results obtained after a second expert delineation are shown.

Table 6: Agreement between our method and the ground truth in terms of left ventricle myocardial mass.

Marker	Volume (cm^3)	ρ	BA Bias
LVM	200.8 (79.0)*	0.899	-24.0 (35.7)

Note: LVM: Left Ventricular Myocardium; ρ: Spearman correlation coefficient; BA: Bland-Altman; * $p < 0.05$ by means of a paired t-Student test.

4.2. Left Ventricular Myocardium Segmentation

The automatic myocardial delineation by means of the proposed U-Net segmentation approach lead to an overall Dice performance of 78.8 ± 7.2% with Haussdorf distance values of 26.2 ± 28.8mm and RVD of −0.08 ± 0.13. As an example of the diverse quality of segmentations achieved, high, middle and low performances obtained at different heart locations are depicted on Fig. 3. The agreement between our method and the ground truth in terms of left ventricle myocardial mass can be appreciated on Table 6.

4.3. Myocardial Abnormality Detection

4.3.1. Model Selection

The herein proposed classification approach was chosen over three classification models explored on a 100 random-splits validation step. Thus, besides the approach explained on Section 3.2.3, we explore a similar one where the features were directly extracted by the non-tuned VGG19 and the remaining algorithm steps (PCA decomposition and SVM fiting) were preserved.

Table 5: Intra- and inter- observers variability results.

Obs	Biomarker	Value	Dice	HD (mm)	RVD	ρ	BA Bias
Intra	$Scar_{vol}$ (cm^3)	38.5 (32.7)*	70.6 (18.3)	17.9 (13.9)	-0.06(0.22)	0.980	-2.2 (7.0)
	LVM_{vol} (cm^3)	191.0(69.1)**	87.5 (2.5)	7.5 (2.3)	-0.09 (0.07)	0.962	-20.0 (18.3)
	IM (%)	17.8 (13.9)*	-	-	-	0.973	-1.0(2.4)
Inter	$Scar_{vol}$ (cm^3)	51.3 (41.5)**	61.2 (24.2)	25.7 (18.7)	0.38 (0.79)	0.915	11.0 (7.04)
	LVM_{vol} (cm^3)	212.9 (77.2)	84.8 (2.8)	8.84 (2.69)	0.01 (0.08)	0.949	1.6 (20.2)
	IM (%)	17.8** (13.9)	-	-	-	0.900	5.2 (9.7)

Note: Obs: Observer; HD: Haussdorf Distance; RVD: Relative Volume Difference; ρ: Spearman Correlation Coefficient; BA: Bland-Altman. LVM: Left Ventricular Myocardium; IM: Infarcted Myocardium. * p < 0.05; ** p <0.01 (paired t-Student tests).

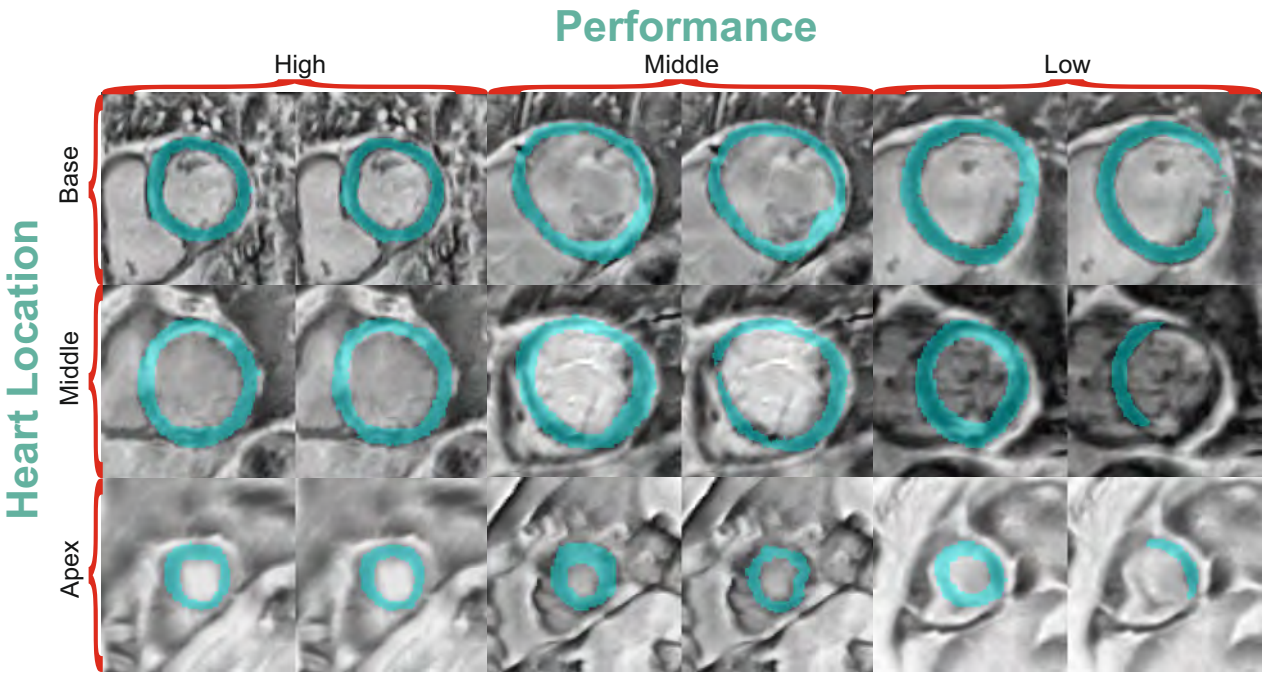


Figure 3: Left ventricle myocardial segmentations obtained. For each image pair, the left (right) one represents the ground truth (obtained result).

Moreover, we also explored a model using direct classification by the fine-tuned VGG19 network. Performance results for the three-proposed models are shown on Fig 4. Overall, the best model selected for classification outperformed the remaining two by achieving a 90.63±4.30% mean accuracy , 88.11±6.53% sensitivity and 93.15±4.83% specificity in a maximum-a-posteriori prediction. The variability for the performance accuracy, sensitivity and specificity metrics were lower for the outperforming model when compared against the other approaches.

4.3.2. ROC Analysis

The AUC value obtained on the ROC analysis was 0.957 ± 0.03 for the proposed classifier, as can be appreciated on Fig 5. The results obtained under 100-random splits scenarios show high performance stability and low variability.

Since the classifier is used in this work to decide whether or not the segmentation lesion search algorithm should be applied in each image, it is not equally important to have false positive or negative detections. Thus, each pathological image misclassified as a healthy one will not be assessed by the segmentation algorithm and their damaged areas will be lost from the analysis. On the other hand, misclassified healthy images into pathological ones might tend to produce an over-segmentation of the lesion. However, the algorithm might handle these images without finding any dam-

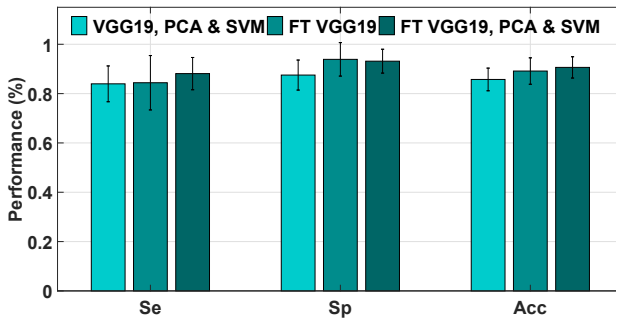


Figure 4: Mean and standard-deviation performance metrics obtained for the 3 explored classification strategies under the 100-random splits validation. Se: Sensitivity; Sp: Specificity; Acc: Accuracy; PCA: Principal Component Analysis; SVM: Support Vector Machines; FT: Fine Tuned.

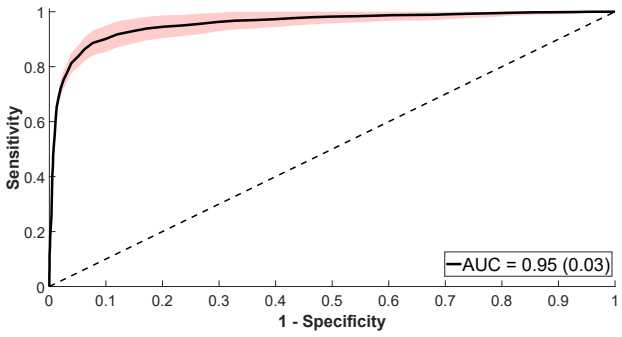


Figure 5: ROC curve obtained after 100-random splits for the proposed classifier. The solid black line represents the mean AUC performance obtained, while the red area represents the variability AUC interval (mean SD). The random-chance classifier is shown with a dashed black line. ROC: Receiver Operating Characteristic; AUC: Area Under the Curve.

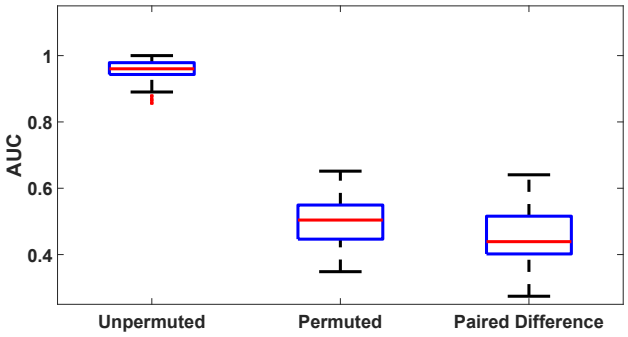


Figure 6: AUC distributions for the permutation analysis.

Table 7: Mean (standard-deviation) segmentation performances obtained for the coarse segmentation followed by different ensemble sizes.

Method	Dice (%)	HD (mm)	RVD
Coarse	73.0 (14.5)	41.89 (17.23)	0.94 (4.07)
Ens #1	76.3 (14.9)	40.90 (17.95)	0.37 (1.75)
Ens #3	76.9 (14.7)	41.36 (17.75)	0.43 (2.07)
Ens #5	77.1 (14.4)	41.31 (17.72)	0.41 (1.90)
Ens #7	**77.2 (14.3)**	**41.25 (17.79)**	**0.41 (1.89)**
Ens #9	77.2 (14.3)	41.28 (17.83)	0.41 (1.93)

Note: Ens: Ensemble Size; Dice: Dice Index; HD: Haussdorf distance; RVD: Relative-volume-difference.

aged area on it. Under this scenario we would like to set up the classifier for assuring high-sensitivity performances. When moving the decision rule threshold for addressing this goal we obtained for sensitivities of 90%, 92.5%, 95% and 97.5% corresponding specificity values of 90%, 85.4%, 73.3% and 57.3%.

The last experiment of this section involved a ROC permutation analysis. Boxplots of AUC permuted, unpermuted and paired differences ($AUC_i^p - AUC_i^{np}$) are shown on Figure 6. It can be appreciated the consistent AUC distribution differences between permuted and un-permuted data, which showed statistical significance ($p < 0.05$, paired t-Student test). The boxplot depicting paired AUC differences shows the abscence of random dataset configurations outperforming in AUC terms the original data configuration.

4.4. Myocardial Scar Segmentation and Quantification

4.4.1. Ensemble Size Selection

The fact of training an ensemble of classifiers using 7 CNNs is based on a comparative analysis conducted for different ensemble sizes. Results obtained for different ensemble models with $k = 1, 3, 5, 7, 9$ components are reported in Table 7. The coarse segmentation by itself achieved an overall Dice index of 73%, which behaved as well or better than all SOA methods (Fig. 7). When the segmentation refinement was introduced, results improved until reaching a mean Dice index of 77.22% for the ensemble using 7 CNNs. It is noticeable that using an ensemble with more CNNs did not improve the segmentation performance. Consequently, after this experiment the amount of CNNSs was fixed on a value of seven and from here on, all presented results are obtained under the chosen configuration.

4.4.2. Segmentation Performance

Achieved segmentation performances for all the compared algorithms are shown on Fig. 7. Our algorithm obtained the highest Dice indexes when compared against the SOA method ones, achieving a Dice value of 77.22 ± 14.3% and considerably outperforming the best ranked SOA method (2-SD with Dice 70.49 ± 16.48%). Besides, our proposal obtained the lowest Dice variance among all methods. Statistical significance was present in all Dice comparisons. When comparing performances in terms of Haussdorf distances, our method obtained 41.2 ± 17.8 mm (Fig 7B). The lowest Haussdorf values were obtained for the 4-SD and 5-SD methods (30.8 ± 16.9 mm and 31.7 ± 18.4 mm respectively, $p < 0.05$). The achieved homocedastic Haussdorf distance distributions showed similar variance levels for all the methods. On Fig. 7C, RVD results are shown. Our proposal achieved a low mean value of 0.4 with high variance. The best RVD performances were obtained by FWHM (-0.19 ± 0.77) and 3-SD methods (-0.24 ± 0.31).

Qualitative segmentation results at different heart locations are shown on Fig. 8. Robustness of the algorithm for detecting the scar at different heart positions can be observed. Overall, less false-positives cluster of pixels were found for our method when comparing against the SOA ones. It is noticeable as well the slight segmentation improvement conducted after the refinement step.

The agreement with the manual delineations obtained by the different methods in terms of clinical markers is summarized on Table 8. Estimation of the scarred myocardial volume as well as of the percentage of infarcted myocardium was consistently better for our proposal when compared against the SOA ones. For both considered metrics, our approach achieved the highest correlation values and lowest Bland-Altman biases. A relevant result is that our proposal was the only method in estimating the scar volume and percentage of infarcted myocardium by agreeing with the manual delineations. All the remaining methods obtained clinical-markers estimations that statistically differed from the expert annotated ones.

4.4.3. Microvascular Obstruction Inclusion

On Table 9 the sensitivity of the different methods for detecting MVO areas are shown. Our proposal achieved the highest performance values and showed statistical

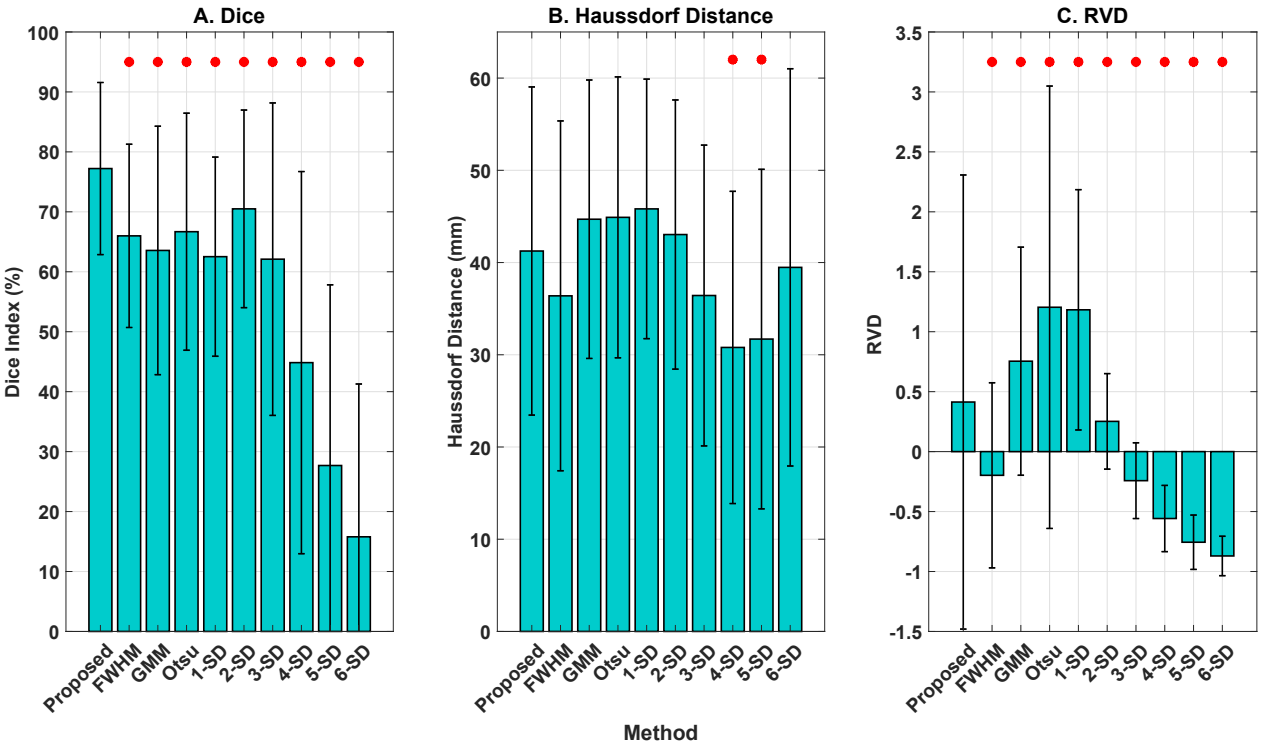


Figure 7: Segmentation performances. RVD: relative-volume-difference; FWHM: Full-width at half-maximum; GMM: Gaussian-mixture-model; n-SD: n-standard deviation thresholding from remote myocardium; $*$: $p < 0.05$ obtained by Mann-Whitney U-test.

Table 8: Agreement between methods and the manual delineations by means of clinical markers.

	Infarcted Volume (cm^3)				Infarcted Myocardium (%)			
Method	Value	ρ	BA Bias	p-value	Value	ρ	BA Bias	p-value
Manual	25.6 (19.3)				18.5 (12.6)			
FWHM	17.5 (12.8)	0.937	-8.0 (8.7)	< 0.001	12.7 (8.1)	0.933	-5.8 (5.9)	< 0.001
GMM	32.7 (17.1)	0.806	7.1 (14.6)	< 0.001	24.1 (10.9)	0.776	5.6 (9.5)	< 0.001
Otsu	39.2 (20.7)	0.906	13.6 (10.0)	< 0.001	28.6 (12.1)	0.884	10.1 (6.5)	< 0.001
1-SD	43.9 (24.3)	0.931	18.3 (10.7)	< 0.001	32.3 (15.7)	0.923	13.7 (7.2)	< 0.001
2-SD	28.6 (18.7)	0.92	2.9 (9.4)	< 0.01	21.1 (13.0)	0.916	2.5 (6.4)	< 0.001
3-SD	18.4 (14.4)	0.874	-7.2 (12.5)	< 0.001	13.6 (10.3)	0.846	-4.8 (7.6)	< 0.001
4-SD	11.2 (11.1)	0.755	-14.4 (15.2)	< 0.001	8.3 (7.9)	0.722	-10.2 (9.4)	< 0.001
5-SD	6.3 (8.4)	0.567	-19.2 (16.9)	< 0.001	4.8 (5.9)	0.565	-13.7 (10.6)	< 0.001
6-SD	3.7 (6.3)	0.458	-21.9 (17.6)	< 0.001	2.7 (4.4)	0.471	-15.7 (11.2)	< 0.001
Proposed	**26.6 (18.5)**	**0.944**	**1.0 (6.8)**	**0.196**	**19.2 (11.0)**	**0.944**	**0.5 (4.5)**	**0.313**

Note: Mean(standard deviation). FWHM: Full-width at half-maximum; GMM: Gaussian-mixture-model; n-SD: n-standard deviation thresholding from remote myocardium; BA: Bland-Altman; ρ: Spearman correlation coefficient; $p - values$ obtained by a paired t-Student test.

Table 9: Mean (standard-deviation) sensitivity for detecting microvascular-obstruction areas per method.

FWHM	GMM	Otsu	1-SD	2-SD	3-SD	4-SD	5-SD	6-SD	**Proposed**
18.8*	54.6*	57.7*	63.6	46.5*	27.9*	14.1*	5.7*	2.8*	**66.9**
(23.0)	(37.1)	(35.0)	(36.4)	(38.9)	(34.0)	(25.1)	(14.6)	(9.1)	**(40.5)**

Note: FWHM: Full-width at half-maximum; GMM: Gaussian-mixture-model; n-SD: n-standard deviation thresholding from remote myocardium. $*$: $p < 0.05$ by means of Mann-Whitney U test.

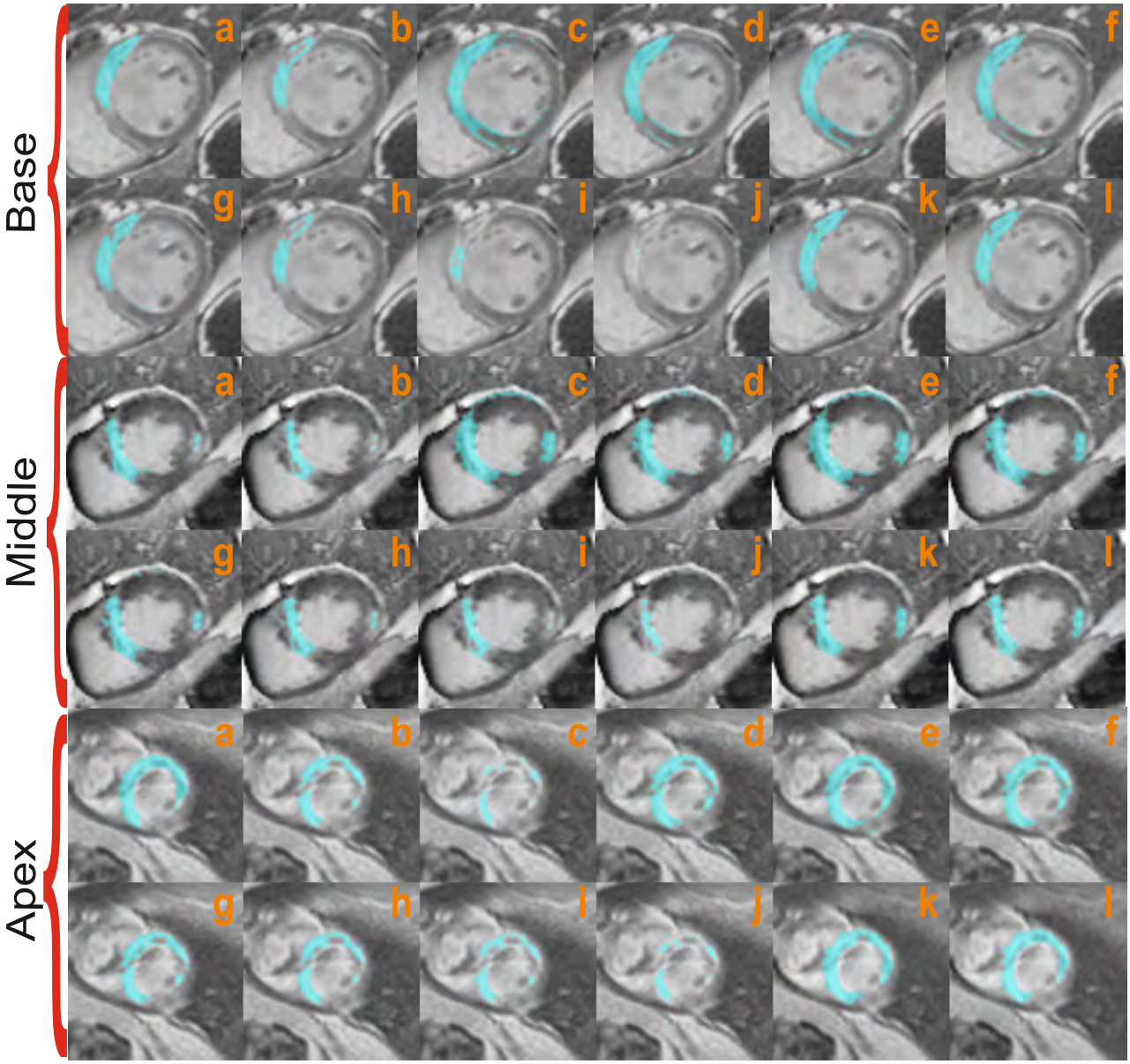


Figure 8: Scar segmentations obtained per algorithm at different heart locations. a) Ground-truth. b) Full-width at half-maximum. c) Gaussian-mixture-models. d) Otsu. e) 1-SD f) 2-SD. g) 3-SD. h) 4-SD. i) 5-SD. j) 6-SD. k) Proposed coarse segmentation. l) Full proposed method.

significance when compared with all SOA methods with exception of the 1-SD one. A MVO segmentation example can be appreciated on Fig. 9, where our proposals capability for the task is exposed. It can be highlighted the accurate segmentation of the hyper-enhanced area provided by the coarse pre-segmentation, with its improvement and MVO inclusion after the refinement approach. For the shown image, only our approach was able to deal with the no-reflow area.

5. Discussion

We present in this work a clinical multi-field, expert delineated reference dataset for cardiac LGE-MRI assessment using phase-sensitive inversion recovery sequences. The large opened dataset will allow direct method comparison and development of machine and deep learning solutions for infarction detection. The dataset accounts with the following main novelties when compared with the only available competing dataset (Karim et al., 2012): *i*) the inclusion of a big cohort of human scans considerably overcoming in size (more than six times greater) the STACOM proposal; *ii*) the full dataset annotation into left ventricular myocardium (labeled as healthy, scarred or MVO) and blood-pool regions. It is worth to highlight the separated delineation of MVO areas which are, to our knowledge, the first time conducted on a dataset

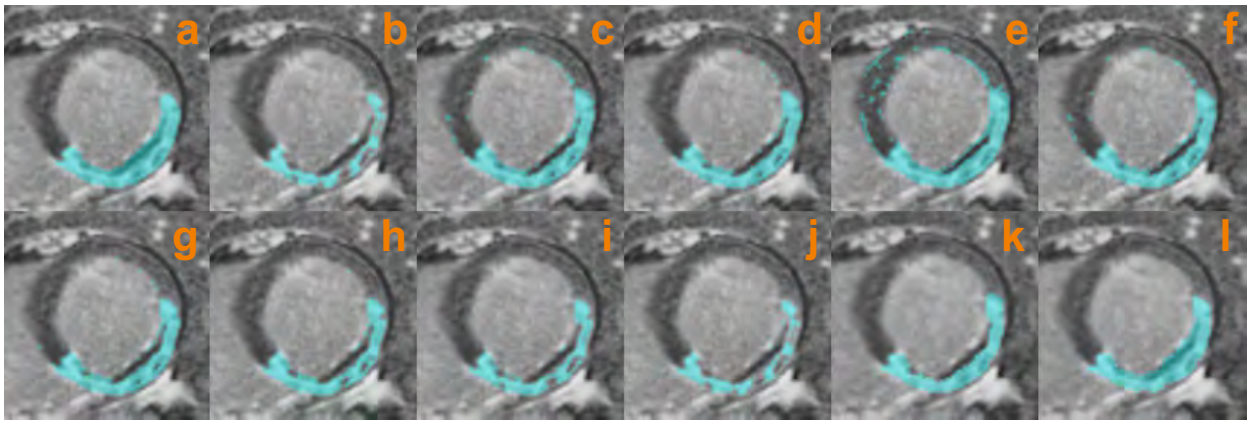


Figure 9: Segmentation results for microvascular-obstructed areas per method. a) Ground-truth. b) Full-width at half-maximum. c) Gaussian-mixture-models. d) Otsu. e) 1-SD f) 2-SD. g) 3-SD. h) 4-SD. i) 5-SD. j) 6-SD. k) Proposed coarse segmentation. l) Full proposed method.

of these characteristics; *iii*) complementary clinical patient information was retrieved and is provided in this work (Table 4), which could allow future deeper analysis and understanding of the myocardial damage phenomena, *iv*) herein it is included a subset (20%) of healthy patients and *v*) different magnetic field scans, which might help in the development of more robust solutions for infarction detection.

When the dataset delineations were assessed by the conduction of intra/inter-observer variability analysis, an overall low agreement was obtained, evidencing the discrepancy between observers for identifying not only the damaged tissue areas, but also the left-ventricle myocardial boundaries. For the intra-observer analysis, despite the high correlation and low Bland-Altman bias obtained for the different clinical markers, all paired analysis evidenced statistically significance with a tendency to under-estimate the myocardium and scar mass (RVD < 0, Table 5). Moreover, the Dice performance was low especially when considering the myocardial scar, with a greater agreement achieved when considering the left ventricular myocardium. On the other hand, the inter-observer analysis showed as expected stronger disagreement in terms of Dice, Spearman correlation and Bland-Altman biases. An exception can be appreciated for the left ventricular myocardium, being the only clinical marker without showing statistical differences with the ground truth. The interpretation of both intra- and inter-observers variability can be attributed to diverse potential causes. Firstly, due to the conduction of fully *blind* delineations. Thus, no previous agreement for conducting the draws was performed between the observers. Although in other medical imaging fields this point could not introduce high discrepancies, the overall small areas to delineate in these studies could be consistently affected. For instance, in the heart base and apex locations where the left ventricular myocardium is highly affected by partial volume effect without exhibiting clearly boundaries, some delineation *rules* should be defined. This is the way that was conducted, for instance, the myocardial ground truth over the ACDC cardiac dataset (Bernard et al., 2018). Secondly, the lack of a consensus for establishing the myocardial clinical condition (healthy or infarcted) for all the patients and slices introduced high variability between observers. It was exhibited that for many images there was disagreement regarding the myocardial clinical status and hence an impact over the clinical markers estimation can be expected. Thirdly, interpretation of Dice metric for evaluating small regions segmentation performance could be critical, specially under the lack of an expert ground truth consensus.

In this work, an end-to-end fully automatic framework for myocardial infarction detection and quantification in LGE-MRI is presented. The strategy is modular and flexible, allowing the use and implementation of not only the entire pipeline, but also of their independent blocks if desired. As well, modifications, improvements, or adaptations for its use in a semi-automatic version (for instance by only using the scar segmentation block after manual delineation) are possible given its block modularity. Among the biggest novelties of the proposed framework we can point out: *i*) The fully user-interactive free capability. Up to know, only Hennemuth et al. (2008) and Wei et al. (2013a) attempted to automatize the whole task. However, their methods showed strong limitations and lack of validation as explained in Section 2. *ii*) The development of a strategy fully independent from Cine MRI sequences. *iii*) The automatic segmentation of the left ventricular myocardium. *iv*) The generalization of the algorithm for working under healthy scenarios as well and *v*) the incorporation on the framework of a dedicated step for including MVO areas within the scar segmentation. All these characteristics turns our framework a highly valuable tool with potential clinical transfer capabilities.

A detailed analysis of the achieved results for the different blocks of the framework is presented below. Our proposal for segmenting the left ventricular myocardium demonstrated an overall good agreement with the expert delineated ground truth. The obtained Dice values were on average just $\sim$ 9% and $\sim$ 6% lesser than the intra- and inter-observer ones, suggesting robustness of our proposal for conducting this task. Moreover, in terms of RVD and Bland-Altman bias our method showed similar performances with the intra-observer obtained ones (Tables 5 and 6), where both delineations exhibited a slight tendency to under-estimate the myocardial mass. Nonetheless, our algorithm's performance was limited in terms of Haussdorf distance and showed a consistent lesser myocardial volume correlation when compared against the intra- and inter- observer cases. For this latter clinical marker, statistically significancy was achieved as happened as well for the intra-observer study.

The qualitative assessment of the myocardial segmentations (Fig. 3) unveal some common limitation patterns for the proposed deep learning approach. Firstly, in most cases the method presented better performances for segmenting the base or middle slices than the apex ones. The reasons might be related with the myocardial size, which is much greater for the base and middle slices than the apex ones, as well as related with the strong image quality differences. On the heart apex, as can be appreciated on Fig. 3, the MRI quality is generally poorer than for central slices, exhibiting non clear boundaries and a strong partial volume effect influence. These results are non surprising and are in agreement with the reported results for kinetic MRI sequences in the ACDC challenge, where the same method's failures were pointed out (Bernard et al., 2018). Secondly, a common limitation of the method was observed for recognizing the myocardial tissue in big infarcted areas. Given that the blood-pool presents similar intensity profiles than the scarred myocardium (Tao et al., 2010), the algorithm presented difficulties in differentiating both regions. It is important to remark that this limitation was mainly observed in big infarctions, since in middle/small scar lesions the remaining myocardial healthy tissue brings contextual information for helping in the task performance. Thirdly, likewise in (Bernard et al., 2018) the method presented unrealistic anatomical configurations for the myocardium, as can be appreciated on Fig. 3. Given that the myocardium geometrically resembles a deformed ring, the inclusion of a-priori information as conducted in Zotti et al. (2017) and/or Ngo et al. (2017) could be a solution for this problem. Despite these limitations, given the hardness of myocardial segmentation in LGE-MRI and the very little solutions published for conducting the task (Table 1) our results turn very promising. To our knowledge, this is the first work presenting a deep learning solution for LGE-MRI myocardium delineation.

By mimicking semi-automatic SOA algorithms pipeline, a classifier was devised for detecting diseased myocardial slices, a step generally conducted by visual inspection from experts. The discriminant rule achieved high classification performance results and was able to overcome the traditional fine tuning approach widely spread in literature. Thus, our results suggest that the fitting of a machine-learning classifier (SVM in our case) feed with fine-tuned CNN features outperforms transfer learning or direct fine-tuned CNN approaches.

When the decision rule was assessed in terms of a ROC analysis, high AUC metrics with low variance were obtained, suggesting robustness of the proposed discriminant rule (Fig. 5). Possible operative points providing high sensitivity were extracted, which will help in reducing the false positive lesions detection in healthy images. Even more, results from the permutation analysis showed that the built classifier and the features used for its devising are informative for the addressed problem and cannot be achieved by a random chance configuration. The permutation analysis excluded the finding of dataset-dependent results and showed that the AUC obtained values do not belong to the null permuted distribution. All these findings supports, consequently, the classifiers robustness as well as the method's reproducibility over different databases.

Segmentation of the infarcted masses was conducted in a two-step approach where the initial segmentation was later improved using deep learning. It is important to highlight the novelty of this approach which was not only thought as a high-performance algorithm, but also as a modular transferable framework. Thus, it is demonstrated the high performance achieved before and after the segmentation refinement, achieving the coarse segmentation step a better agreement with the ground truth than the SOA methods ($\sim$ 3% greater Dice than the remaining algorithms). Given its easiness for implementation without requiring GPU computation capabilities, the coarse segmentation algorithm by itself could be used in clinical practice as well.

When the deep-learning based refinement was included, a consistent and statistical significant improvement in segmentation agreement was achieved ($\sim$7% greater Dice than the best performing SOA). Even more, the low Dice variance showed homogeneity and adaptability of the method to different myocardial lesions configurations. However, when assessing the Haussdorf distance results, our method obtained non-outstanding performances. Overall, it performed similar to most SOA algorithms excepting 4-SD and 5-SD, which achieved much lower metrics ($p < 0.05$). These results are expectable since in these algorithms the segmentation histogram threshold is set very high. Thus, only highly hyperenhanced voxels belonging to the core necrotic tissue are detected and false detections

coming from overlapped histogram areas are avoided. Given the fact that the Haussdorf distance is strongly affected by outliers (Jia et al., 2018), these methods result benefited by this metric.

When comparing RVD metric performances, the proposed method achieved a mean low positive value suggesting a tendency to slightly over-segment the lesions. Nevertheless, a high variance can be appreciated evidencing the presence of not only over-segmented but under-segmented lesions as well. The best RVD performances were obtained for FWHM and 3-SD methods, whose results under-estimated the lesions mass. Despite the good performances obtained by 3-4-5-SD methods in terms of HD and RVD values, very poor Dice agreement with the ground truth characterized these algorithms without suggesting suitability for clinical usage (Fig. 7).

Promising results in terms of clinical markers were achieved with the proposed algorithm. The high correlation, low bias and the fact of being the only method agreeing in volumetric lesion quantification with the delineations (Table 8) suggest its appropriateness for working under clinical and medical scenarios. On the other hand, supporting the findings of Spiewak et al. (2010) and Zhang et al. (2016) the SOA results showed very poor scar segmentation agreement with the manual delineations, characterized by low accuracies, high results variability and significant differences in volumetric tissue quantification.

Considering the novelty of the used dataset that contains no-reflow annotated cases, the inclusion of MVO areas within the infarcted segmented masks was compared between the different methods. Our approach was consistently superior for conducting this task, achieving the highest sensitivity performance and evidencing statistical significance when compared against the SOA approaches. The 1-SD method was the only exception, showing non-significant differences even when achieving lower performances (Table 9). For this latter technique, the setting of a very low threshold for detecting myocardial scars favors MVO detection at the expenses of providing low overall performances (Fig. 7 and Table 8).

6. Limitations and Future Work

The main dataset limitation regards the reliability of the expert contours. For decreasing the intra- and inter-observer delineation variabilities, a consensus between observers will be established. It will include as well the definition of a protocole for conducting the annotations, specially for the basal and apical slices as conducted for the ACDC challenge (Bernard et al., 2018). With respect to the proposed framework, the main limitation regards the optimal conditions results reported in this study. In forthcoming analysis the impact of the different block algorithms over the final scar quantification will be conducted.

7. Conclusions

The findings of this study can be summarized into two main contributions. Firstly, herein is proposed for the very first time a big multifield open dataset for myocardial infarction quantification in LGE-MRI. It includes annotated data from the left ventricular myocardium, hyper-enhanced and no-reflow regions. To our knowledge, this is the first dataset including several cases with no-reflow phenomena and were its regions are expert annotated. Secondly, we propose and validate an end-to-end fully automatic framework for infarction segmentation and quantification. The framework overcomes several limitations of previous proposals from which can be highlighted: *i*)The only use of LGE-MRI data and its independence from cine MRI, *ii*) the automatic segmentation of the left ventricular myocardium, *iii*) the detection of disease images, allowing to extend the method for working under healthy scenarios and *iv*) the development of a novel and robust technique for automatically delineating the scar tissue. The extensive statistical validation of the framework and its vast comparison against several current state-of-the-art methods turn this proposal into a robust and reliable tool with clinical transfer potential.

8. Acknowledgments

Ezequiel de la Rosa holds an Erasmus Mundus master's scholarship. The authors would like to thank Dr. Thibault Leclercq for helping in the dataset delineation and NVIDIA Corporation for the donation of the TITAN-X GPU used in this reseach.

References

Amado, L.C., Gerber, B.L., Gupta, S.N., Rettmann, D.W., Szarf, G., Schock, R., Nasir, K., Kraitchman, D.L., Lima, J.A., 2004. Accurate and objective infarct sizing by contrast-enhanced magnetic resonance imaging in a canine myocardial infarction model. Journal of the American College of Cardiology 44, 2383–2389.

Andreu, D., Berruezo, A., Ortiz-Pérez, J.T., Silva, E., Mont, L., Borràs, R., de Caralt, T.M., Perea, R.J., Fernández-Armenta, J., Zeljko, H., et al., 2011. Integration of 3d electroanatomic maps and magnetic resonance scar characterization into the navigation system to guide ventricular tachycardia ablationclinical perspective. Circulation: Arrhythmia and Electrophysiology 4, 674–683.

Antropova, N., Huynh, B.Q., Giger, M.L., 2017. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. Medical physics .

Arai, A.E., 2008. Myocardial infarction and viability with an emphasis on imaging delayed enhancement, in: Cardiovascular Magnetic Resonance Imaging. Springer, pp. 351–375.

Avendi, M., Kheradvar, A., Jafarkhani, H., 2016. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac mri. Medical image analysis 30, 108–119.

BahadarKhan, K., Khaliq, A.A., Shahid, M., 2016. A morphological hessian based approach for retinal blood vessels segmentation and denoising using region based otsu thresholding. PloS one 11, e0158996.

Baumgartner, C.F., Koch, L.M., Pollefeys, M., Konukoglu, E., 2017. An exploration of 2d and 3d deep learning techniques for cardiac mr image segmentation. arXiv preprint arXiv:1709.04496 .

Bernal, J., Kushibar, K., Cabezas, M., Valverde, S., Oliver, A., Lladó, X., 2018. Quantitative analysis of patch-based fully convolutional neural networks for tissue segmentation on brain magnetic resonance imaging. arXiv preprint arXiv:1801.06457 .

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al., 2018. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Transactions on Medical Imaging .

Bland, J.M., Altman, D., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. The lancet 327, 307–310.

Carminati, M.C., Boniotti, C., Fusini, L., Andreini, D., Pontone, G., Pepi, M., Caiani, E.G., 2016. Comparison of image processing techniques for nonviable tissue quantification in late gadolinium enhancement cardiac magnetic resonance images. Journal of thoracic imaging 31, 168–176.

Ciofolo, C., Fradkin, M., Mory, B., Hautvast, G., Breeuwer, M., 2008. Automatic myocardium segmentation in late-enhancement mri, in: Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on, IEEE. pp. 225–228.

Curiale, A.H., Colavecchia, F.D., Kaluza, P., Isoardi, R.A., Mato, G., 2017. Automatic myocardial segmentation by using a deep learning network in cardiac mri. arXiv preprint arXiv:1708.07452 .

Dastidar, A.G., Rodrigues, J.C., Baritussio, A., Bucciarelli-Ducci, C., 2015. Mri in the assessment of ischaemic heart disease. Heart , heartjnl–2014.

Detsky, J.S., Paul, G., Dick, A.J., Wright, G.A., 2009. Reproducible classification of infarct heterogeneity using fuzzy clustering on multicontrast delayed enhancement magnetic resonance images. IEEE transactions on medical imaging 28, 1606–1614.

Dikici, E., ODonnell, T., Setser, R., White, R.D., 2004. Quantification of delayed enhancement mr images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 250–257.

Durante, A., Camici, P.G., 2015. Novel insights into an old phenomenon: the no reflow. International journal of cardiology 187, 273–280.

Engblom, H., Tufvesson, J., Jablonowski, R., Carlsson, M., Aletras, A.H., Hoffmann, P., Jacquier, A., Kober, F., Metzler, B., Erlinge, D., et al., 2016. A new automatic algorithm for quantification of myocardial infarction imaged by late gadolinium enhancement cardiovascular magnetic resonance: experimental validation and comparison to expert delineations in multi-center, multi-vendor patient data. Journal of Cardiovascular Magnetic Resonance 18, 27.

Ernst, M.D., et al., 2004. Permutation methods: a basis for exact inference. Statistical Science 19, 676–685.

Flett, A.S., Hasleton, J., Cook, C., Hausenloy, D., Quarta, G., Ariti, C., Muthurangu, V., Moon, J.C., 2011. Evaluation of techniques for the quantification of myocardial scar of differing etiology using cardiac magnetic resonance. JACC: cardiovascular imaging 4, 150–156.

Ginat, D.T., Fong, M.W., Tuttle, D.J., Hobbs, S.K., Vyas, R.C., 2011. Cardiac imaging: part 1, mr pulse sequences, imaging planes, and basic anatomy. American Journal of Roentgenology 197, 808–815.

Guerrero, R., Qin, C., Oktay, O., Bowles, C., Chen, L., Joules, R., Wolz, R., Valdés-Hernández, M., Dickie, D., Wardlaw, J., et al., 2018. White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. NeuroImage: Clinical 17, 918–934.

Hennemuth, A., Seeger, A., Friman, O., Miller, S., Klumpp, B., Oeltze, S., Peitgen, H.O., 2008. A comprehensive approach to the analysis of contrast enhanced cardiac mr images. IEEE Transactions on Medical Imaging 27, 1592–1610.

Hsu, L.Y., Natanzon, A., Kellman, P., Hirsch, G.A., Aletras, A.H., Arai, A.E., 2006. Quantitative myocardial infarction on delayed enhancement mri. part i: Animal validation of an automated feature analysis and combined thresholding infarct sizing algorithm. Journal of Magnetic Resonance Imaging 23, 298–308.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 .

Isensee, F., Jaeger, P., Full, P.M., Wolf, I., Engelhardt, S., Maier-Hein, K.H., 2017. Automatic cardiac disease assessment on cine-mri via time-series segmentation and domain specific features. arXiv preprint arXiv:1707.00587 .

Jia, H., Xia, Y., Song, Y., Cai, W., Fulham, M., Feng, D.D., 2018. Atlas registration and ensemble deep convolutional neural network-based prostate segmentation using magnetic resonance imaging. Neurocomputing 275, 1358–1369.

Karim, R., Bhagirath, P., Claus, P., Housden, R.J., Chen, Z., Karimaghaloo, Z., Sohn, H.M., Rodríguez, L.L., Vera, S., Albà, X., et al., 2016. Evaluation of state-of-the-art segmentation algorithms for left ventricle infarct from late gadolinium enhancement mr images. Medical image analysis 30, 95–107.

Karim, R., Claus, P., Chen, Z., Housden, R.J., Obom, S., Gill, H., Ma, Y., Acheampong, P., ONeill, M., Razavi, R., et al., 2012. Infarct segmentation challenge on delayed enhancement mri of the left ventricle, in: International Workshop on Statistical Atlases and Computational Models of the Heart, Springer. pp. 97–104.

Kim, R.J., Fieno, D.S., Parrish, T.B., Harris, K., Chen, E.L., Simonetti, O., Bundy, J., Finn, J.P., Klocke, F.J., Judd, R.M., 1999. Relationship of mri delayed contrast enhancement to irreversible injury, infarct age, and contractile function. Circulation 100, 1992–2002.

Kolipaka, A., Chatzimavroudis, G.P., White, R.D., ODonnell, T.P., Setser, R.M., 2005. Segmentation of non-viable myocardium in delayed enhancement magnetic resonance images. The international journal of cardiovascular imaging 21, 303–311.

Kotu, L.P., Engan, K., Skretting, K., Ørn, S., Woie, L., Eftestøl, T., 2013. Segmentation of scarred myocardium in cardiac magnetic resonance images. ISRN Biomedical Imaging 2013.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

Kruk, D., Boucher, A., Lalande, A., Cochet, A., Sliwa, T., 2017. Segmentation integrating watershed and shape priors applied to cardiac delayed enhancement mr images. IRBM 38, 224–227.

Kurzendorfer, T., Forman, C., Brost, A., Maier, A., 2017a. Random forest based left ventricle segmentation in lge-mri, in: International Conference on Functional Imaging and Modeling of the Heart, Springer. pp. 152–160.

Kurzendorfer, T., Forman, C., Schmidt, M., Tillmanns, C., Maier, A., Brost, A., 2017b. Fully automatic segmentation of left ventricular anatomy in 3-d lge-mri. Computerized Medical Imaging and Graphics 59, 13–27.

Lalande, A., Garreau, M., Frouin, F., et al., 2015. Evaluation of cardiac structure segmentation in cine magnetic resonance imaging. Multi-Modality Cardiac Imaging , 169–215.

Larroza, A., López-Lereu, M.P., Monmeneu, J.V., Bodí, V., Moratal, D., 2017. Texture analysis for infarcted myocardium detection on delayed enhancement mri, in: Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on, IEEE. pp. 1066–1069.

Liu, J., Zhuang, X., Wu, L., An, D., Xu, J., Peters, T., Gu, L., 2017. Myocardium segmentation from de mri using multicomponent gaussian mixture model and coupled level set. IEEE Transactions on Biomedical Engineering 64, 2650–2661.

Lu, Y., Yang, Y., Connelly, K.A., Wright, G.A., Radau, P.E., 2012. Automated quantification of myocardial infarction using graph cuts on contrast delayed enhanced magnetic resonance images. Quantitative imaging in medicine and surgery 2, 81.

Manjón, J.V., Coupé, P., Martí-Bonmatí, L., Collins, D.L., Robles, M., 2010. Adaptive non-local means denoising of mr images with spatially varying noise levels. Journal of Magnetic Resonance Imaging 31, 192–203.

Ngo, T.A., Lu, Z., Carneiro, G., 2017. Combining deep learning and level set for the automated segmentation of the left ventricle of the heart from cardiac cine magnetic resonance. Medical image analysis 35, 159–171.

O'Donnell, T.P., Xu, N., Setser, R.M., White, R.D., 2003. Semi-automatic segmentation of nonviable cardiac tissue using cine and delayed enhancement magnetic resonance images, in: Medical Imaging 2003: Physiology and Function: Methods, Systems, and Applications, International Society for Optics and Photonics. pp. 242–252.

Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., Cook, S.A., de Marvao, A., Dawes, T., ORegan, D.P., et al., 2018. Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation. IEEE transactions on medical imaging 37, 384–395.

Organization, W.H., Organization, W.H., et al., 2011. The atlas of heart disease and stroke. 2004. World Health Organization: Geneva .

Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics 9, 62–66.

Pattanayak, P., Bleumke, D.A., 2015. Tissue characterization of the myocardium: state of the art characterization by magnetic resonance and computed tomography imaging. Radiologic Clinics 53, 413–423.

Petitjean, C., Dacher, J.N., 2011. A review of segmentation methods in short axis cardiac mr images. Medical image analysis 15, 169–184.

Pop, M., Ghugre, N.R., Ramanan, V., Morikawa, L., Stanisz, G., Dick, A.J., Wright, G.A., 2013. Quantification of fibrosis in infarcted swine hearts by ex vivo late gadolinium-enhancement and diffusion-weighted mri methods. Physics in Medicine & Biology 58, 5009.

Positano, V., Pingitore, A., Giorgetti, A., Favilli, B., Santarelli, M.F., Landini, L., Marzullo, P., Lombardi, M., 2005. A fast and effective method to assess myocardial necrosis by means of contrast magnetic resonance imaging. Journal of Cardiovascular Magnetic Resonance 7, 487–494.

Rajchl, M., Yuan, J., White, J.A., Ukwatta, E., Stirrat, J., Nambakhsh, C.M., Li, F.P., Peters, T.M., 2014. Interactive hierarchical-flow segmentation of scar tissue from late-enhancement cardiac mr images. IEEE Transactions on Medical Imaging 33, 159–172.

Rajiah, P., Desai, M.Y., Kwon, D., Flamm, S.D., 2013. Mr imaging of myocardial infarction. Radiographics 33, 1383–1412.

Ram, S., Rodríguez, J.J., Bosco, G., 2012. Segmentation and detection of fluorescent 3d spots. Cytometry Part A 81, 198–212.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Savelli, B., Marchesi, A., Bria, A., Marrocco, C., Molinara, M., Tortorella, F., 2017. Retinal vessel segmentation through denoising and mathematical morphology, in: International Conference on Image Analysis and Processing, Springer. pp. 267–276.

Schmidt, A., Azevedo, C.F., Cheng, A., Gupta, S.N., Bluemke, D.A., Foo, T.K., Gerstenblith, G., Weiss, R.G., Marbán, E., Tomaselli, G.F., et al., 2007. Infarct tissue heterogeneity by magnetic resonance imaging identifies enhanced cardiac arrhythmia susceptibility in patients with left ventricular dysfunction. Circulation 115, 2006–2014.

Sidibe, D., Sankar, S., Lemaitre, G., Rastgoo, M., Massich, J., Cheung, C.Y., Tan, G.S., Milea, D., Lamoureux, E., Wong, T.Y., et al., 2017. An anomaly detection approach for the identification of dme patients using spectral domain optical coherence tomography images. Computer methods and programs in biomedicine 139, 109–117.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Spiewak, M., Malek, L.A., Misko, J., Chojnowska, L., Milosz, B., Klopotowski, M., Petryka, J., Dabrowski, M., Kepka, C., Ruzyllo, W., 2010. Comparison of different quantification methods of late gadolinium enhancement in patients with hypertrophic cardiomyopathy. European journal of radiology 74, e149–e153.

Tao, Q., Milles, J., Zeppenfeld, K., Lamb, H.J., Bax, J.J., Reiber, J.H., van der Geest, R.J., 2010. Automated segmentation of myocardial scar in late enhancement mri using combined intensity and spatial information. Magnetic Resonance in Medicine 64, 586–594.

Ukwatta, E., Arevalo, H., Li, K., Yuan, J., Qiu, W., Malamas, P., Wu, K.C., Trayanova, N.A., Vadakkumpadan, F., 2016. Myocardial infarct segmentation from magnetic resonance images for personalized modeling of cardiac electrophysiology. IEEE transactions on medical imaging 35, 1408–1419.

Valindria, V.V., Angue, M., Vignon, N., Walker, P.M., Cochet, A., Lalande, A., 2011. Automatic quantification of myocardial infarction from delayed enhancement mri, in: Signal-Image Technology and Internet-Based Systems (SITIS), 2011 Seventh International Conference on, IEEE. pp. 277–283.

Wei, D., Sun, Y., Ong, S.H., Chai, P., Teo, L.L., Low, A.F., 2013a. A comprehensive 3-d framework for automatic quantification of late gadolinium enhanced cardiac magnetic resonance images. IEEE Transactions on Biomedical Engineering 60, 1499–1508.

Wei, D., Sun, Y., Ong, S.H., Chai, P., Teo, L.L., Low, A.F., 2013b. Three-dimensional segmentation of the left ventricle in late gadolinium enhanced mr images of chronic infarction combining long-and short-axis information. Medical image analysis 17, 685–697.

Xu, C., Xu, L., Gao, Z., Zhao, S., Zhang, H., Zhang, Y., Du, X., Zhao, S., Ghista, D., Li, S., 2017. Direct detection of pixel-level myocardial infarction areas via a deep-learning algorithm, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 240–249.

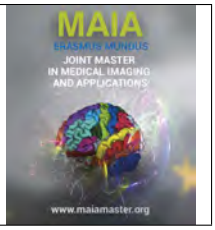
Zhang, L., Huttin, O., Marie, P.y., Felblinger, J., Beaumont, M., CHILLOU, C.D., Girerd, N., Mandry, D., 2016. Myocardial infarct sizing by late gadolinium-enhanced mri: Comparison of manual, full-width at half-maximum, and n-standard deviation methods. Journal of Magnetic Resonance Imaging 44, 1206–1217.

Zotti, C., Luo, Z., Lalande, A., Humbert, O., Jodoin, P.M., 2017. Gridnet with automatic shape prior registration for automatic mri cardiac segmentation. arXiv preprint arXiv:1705.08943 .

Zreik, M., Lessmann, N., van Hamersvelt, R.W., Wolterink, J.M., Voskuil, M., Viergever, M.A., Leiner, T., Išgum, I., 2018. Deep learning analysis of the myocardium in coronary ct angiography for identification of patients with functionally significant coronary artery stenosis. Medical image analysis 44, 72–85.

Medical Imaging and Applications

Master Thesis, June 2018



Improving Breast Cancer Detection using Symmetry Information with Deep Learning

Yeman Brhane Hagos[a,b,c], Albert Gubern Mérida[b], Jonas Teuwen[b]

[a] *University of Burgundy (France), UNICLAM (Italy) and University of Girona (Spain)*
[b] *Radboud University Medical Center, Department of Radiology and Nuclear Medicine, Nijmegen, the Netherlands*
[c] *yemanbrhane1989@gmail.com*

Abstract

Breast cancer is the second dominant cause of cancer death among women, and there has been a significant amount of research to develop Computer-aided detection (CAD) systems for early stage detection and diagnosis. Although, Convolutional Neural Networks (CNN) has had a huge success in many areas of computer vision and medical image analysis, in mammogram breast cancer detection CAD systems there is an immense potential of performance improvement by integrating all the information that radiologist utilize, such as symmetry and temporal data. In this work, to integrate symmetry information into CNN, we propose a patch based multi-input CNN that learns symmetrical difference to detect malignant breast mass in Digital Mammogram (DM) images. First, a candidate detector that uses local lines and gradient orientation based features are employed. Then, a CNN that incorporates symmetry information was adopted to reduce the False Positive (FP) with high sensitivity. To alleviate imbalance between pathological observations and normal candidates an efficient augmentation with perturbation was applied. The network is trained on a large-scale dataset of 28294 DM images collected from different sites and obtained using machines from three different vendors. A baseline architecture without symmetry information was also trained. We observed that integrating symmetrical information slightly outperforms the baseline architecture. Performance of the baseline architecture and symmetry CNN were evaluated using Area Under the ROC Curve (AUC) and Competition Performance Metric (CPM) based average Free Receiver Operating Characteristic (FROC) sensitivity. At candidate level, AUC value of 0.933 with 95% confidence interval of [0.920 , 0.954] was obtained when symmetry information is incorporated in comparison with baseline architecture which yielded AUC value of 0.929 with [0.919 , 0.947] confidence interval. Although there was no a significant candidate level performance again (p = 0.111) of incorporating symmetrical information, we have found a compelling result at exam level with CPM value of 0.733 (p = 0.001). We believe that including temporal data, and adding benign class to the dataset could improve the detection performance.

Keywords: Breast cancer, Digital mammogram, Convolutional neural networks, Symmetry, Deep learning

1. Introduction

Breast cancer is the second most cancer causing death in women after lung cancer in the United States and the chance of a woman dying from breast cancer is 2.6% Siegel et al. (2018), which covers around 30% of cancers diagnosed (Rakhlin et al., 2018). According to American Cancer Society in 2018, around 266120 and 63960 new cases of carcinoma and invasive carcinoma breast cancer will be diagnosed in women, respectively. Approximately 40920 women will die from breast cancer.

Breast cancer is a complex disease in which its etiology is not fully understood due to its multifactorial nature. This makes its prognosis difficult, however, studies have revealed that its main risk factors include genetics and hormones Martin and Weber (2000), and environment, sociobiological (age and gender) and physiological factors also affect its development (Organization et al., 2006). Women in different geographical location have shown a different potential of developing breast cancer. Stefan (2015) stated that in Africa, presumably around 35 per 100,000 women in most countries

(as compared to over 90 – 120 per 100,000 women in most European or North American countries) will have breast cancer. Signs of breast cancer may include a lump in the breast, a change in breast shape, dimpling of the skin, fluid coming from the nipple, a newly inverted nipple, or a red or scaly patch of skin. In those with distant spread of the disease, there may be bone pain, swollen lymph nodes, shortness of breath, or yellow skin (Organization et al., 2006).

For a better survival, early detection and improved diagnosis of breast cancer are essential. There are different breast imaging modalities. Some of them are used for screening purpose, others for diagnosis. Once breast cancer has been detected in screening stage, more detailed evaluations are usually performed using diagnostic modalities which may also be used for initial diagnosis.

The currently used modalities include mammography, breast ultrasound, Tomosynthesis, magnetic resonance imaging (MRI), positron emission tomography (PET), and computed tomography (CT). Mammography is the most common method of breast imaging. It uses low-dose amplitude X-rays to examine the human breast. Cancerous masses and calcium deposits appear brighter on the mammogram. This method is good for detecting Ductal Carcinoma In Situ (DCIS) and calcification (Sree et al., 2011).

Tomosynthesis produces a 3-dimensional image of the breast acquired using several low dose x-rays from different angles. It allows the radiologist to see through dense tissue to a greater degree than with 2-dimensional mammography. This technique improves the chances of finding cancer early and reduces the risk of a false alarm (Gilbert et al., 2016).

Breast cancer screening is important because, for some breast cancers, symptoms might not be visible until cancer reaches a certain level. During breast cancer screening, radiologists utilize Medio Lateral Oblique (MLO) and Cranio-Caudal (CC) views of mammography scan of both breasts to identify markers of lesion such as masses and calcification (Bick, 2014), (Giger et al., 2013). The radiologist looks for the shape and appearance of some suspicious regions and characterizes them as malignant and normal. Breast masses are most dense and appear in grey to white pixel intensity with oval or irregular shape while micro-calcification is characterized by a cluster of small round bright spots in the breast (Oliver et al., 2010), (Tang et al., 2009). Normally, irregular or spicule shaped masses are considered as malignant and malignancy of micro-calcification depends on the location of the marker in the breast (Oliver et al., 2010), (Dhungel et al., 2017).

In most countries, women between the age of 45 and 65 are recommended for breast cancer screening at a regular interval depending on the country where they live. This has shown a reduction in mortality rate between 40% and 45% for women who were undergoing mammogram screening (Feig, 2002), (Group et al., 2006). Despite this general benefits, mammogram screening has harms associated with FP recalls which results in FP biopsy and anxiety caused by the recall for additional diagnostic test after screening (Tosteson et al., 2014), (Oeffinger et al., 2015). There is also a cost associated with these unnecessary follow up (Geras et al., 2017). Therefore, it is necessary to increase sensitivity for early stage detection and increase specificity to reduce FP.

Moreover, the large volume of mammogram images makes the manual screening tedious task for radiologists Dhungel et al. (2017), Anttinen et al. (1993) has stated manual screening has low sensitivity and high recall rate.

Nowadays, with a massive amount of data and computational power, Deep Learning (DL) has showed a remarkable success in the natural language Bahdanau et al. (2014), Iyyer et al. (2015), and object detection and recognition (Wang et al., 2016b). This has opened an interest in applying DL in medical image processing and analysis and it has shown a potential improvement in detection and classification problems. However, care should be taken as the way we humans interpret natural images and medical images are different, for example, in mammogram breast cancer detection and diagnosis. Eventually, the performance of DL method will be compared with the radiologist and the network should be given all the information that radiologist use. For instance, during the reading of screening mammograms, radiologists visually compare the latest images (current) to the ones acquired in the previous screening round (prior). Similarly, right and left breasts are also compared as shown in figure 1 in order to find differences in the breast tissue (asymmetries) that might be signs of cancer. This information has been found to be very important to improve overall detection performance and reduce false positive recalls. Moreover, CC and MLO views are considered.

In object detection and recognition problems, images are often down-scaled while keeping the performance at the same level, however, in medical applications, fine details are needed in detection, classification, and segmentation. Original resolution is desirable for early-stage breast cancer detection. Geras et al. (2017) proposed a multi-view single stage CNN that works at original resolution to classify MLO and CC views. To address memory issue, they proposed aggressive convolution and pooling layers with stride greater than one instead of downscaling the image beforehand. There is spatial information loss, though it is better than downscaling in the image space. Adopting a patch based approach could address this problem. Incorporating symmetry and temporal information improves detection of malignant soft tissue lesion, in which random forest classifier was used for mass detection and CNN for classification Kooi and Karssemeijer (2017).

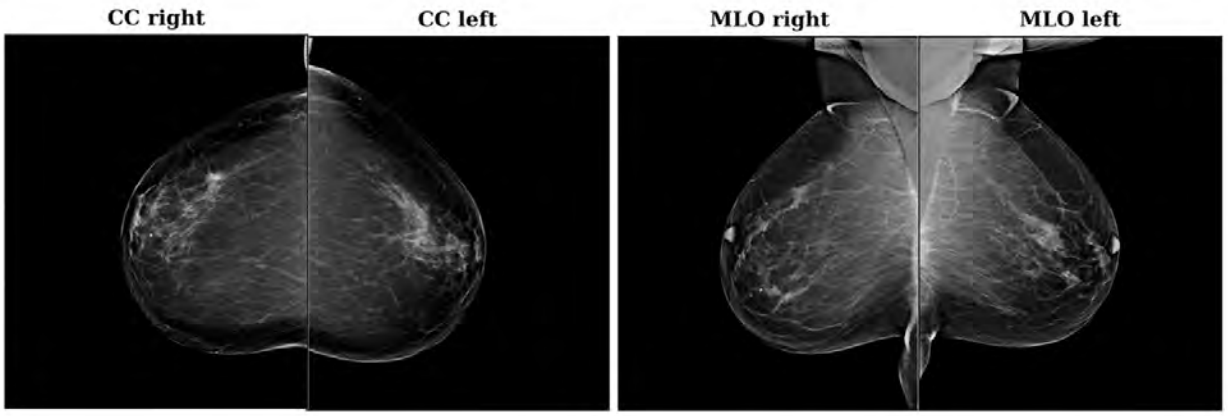


Figure 1: An illustrative example showing symmetry and different views of left and right breast

In this study, we conducted an investigation to analyze the performance gain of integrating symmetry information to CNN to detecting malignant lesions on a large scale mammography images. First, a database of 7196 exams which contains 28294 images was collected from different sites in the Netherlands. Previous work of (Karssemeijer, 1999) which has high sensitivity was used to detect suspicious candidates. Then, patches centered on the points were extracted to train a two input CNN to reduce FP candidates. Left and right breast images were considered as contra-later images to each other and a patch in a primary image and an exact reflection or mirror on the contra-lateral were considered as a pair of inputs. In the work of Kooi et al. (2017) incorporating both symmetry and temporal context was investigated. In this study, we investigated the performance gain of incorporating symmetry context only.

One of the drawbacks of two-stage decision support system is that it produces sub-optimal results as the stages are optimized independently (Dhungel et al., 2017). To overcome this limitation the first stage used in this study has h high sensitivity and low specificity. Moreover, an efficient and effective way of augmentation to train a CNN using imbalanced data is presented which can be used in other medical image analysis problems. Furthermore, qualitative, and quantitative analysis of the proposed method is presented.

This paper is organized as follows. In section two, state of art papers is reviewed. The dataset used and methodology is described in section three. Then, results and statistical analysis are presented in section four. Finally, discussion and conclusion are presented in section five and six, respectively.

2. State of the Art

Several works have used deep learning to perform mammogram lesion classification and detection. Table 1 summaries the current state of art deep learning based approaches to breast mammography. The table presents the target of the study, the size of the dataset, image size, performance evaluation metric, number of stages and type of input image considered.

In mammography, the most frequently used preprocessing techniques are intensity normalization and region of interest (ROI) extraction. In multi-stage breast cancer screening or detection, mammogram image is divided into different regions or suspicious candidates are detected either handcrafted features Wang et al. (2016a). Then, in the next stage, each of these regions will be processed to predict its class. There are also research works that process directly the whole image. As can be seen in the Table 1, the most frequently used metric is AUC.

As the size of mammogram images is too large to fit in a Graphics Processing Unit (GPU), most of the recent researches have focused on multi-stage. For instance, mass detection Kooi and Karssemeijer (2017), Domingues and Cardoso (2013), Carneiro and Bradley, Ertosun and Rubin (2015), Lévy and Jain (2016), Jiao et al. (2016) and micro-classification Mordang et al. (2016), lesion level detection (Lévy and Jain (2016), Arevalo et al. (2016)).

Kooi and Karssemeijer (2017) and Carneiro and Bradley employed random forest classifier for candidate detection. Then, patches were extracted to train a CNN (patch size can be seen in the table). In the work of Kooi and Karssemeijer (2017) in addition to the main patch, contra-lateral and temporal information was added and it is stated that incorporating these additional information increases the performance of the classifier. A large-scale dataset of 73464 images was used and negative candidates were taken only from exams without lesion.

In the work of Jiao et al. (2016), and Arevalo et al. (2016), Huynh et al. (2016) Support Vector Machine (SVM) classifier was trained on features extracted from CNN. Others have focused on training a CNN for classifying small region of interest in to either malignant or benign lesion(Huynh et al. (2016), Mordang et al. (2016), Lévy and Jain (2016)).

Alternatively, a small number of researchers have considered training the whole image and replaced the multi-stage by single stage which can be trained end-to-end. Geras et al. (2017) has proposed high-resolution breast cancer screening with multi-view CNN and evaluated on a large scale data of around 890 thousand images. In contrast to natural images, in medical image analysis original resolution is necessary to detect abnormalities like masses and calcification at an early stage and downsampling hides original patterns that are determinant. However, training on the original resolution of mammogram images is limited by memory requirement. To address this issue, Geras et al. (2017) has proposed aggressive convolution and pooling layers with stride greater than one in the earlier layers and work on the original resolution.

However, Geras et al. (2017) has stated also aggressive convolution has a drawback of loss of spatial information. This will have less effect in classification, but in detection, it will affect negatively as localization is important. Training with images at original resolution and increasing training data increases classifier performanceGeras et al. (2017). A multi-view deep residual network was also proposed by Dhungel et al. (2017) to classify breast mammogram image into malignant and benign. Both views and binary mask of masses and micro-calcification for each view were given as an input to an ensemble of deep residual networks. Detection of lesion and segmentation was done using (Carneiro and Bradley, Lu et al. (2016)). The main drawback of this implementation was the number of inputs. For one breast CC and MLO view images and additional 4 binary mask image are generated and fed as input to the network.

3. Material and Methods

3.1. Dataset

The mammogram images used were collected from General Electric, Siemens, and Hologic from women attending for diagnostic purpose between 2000 and 2016. The images are anonymized and approved by the regional ethics board after summary review, with a waiver of a full review and informed consent (de Moor et al., 2018). The database contains 7196 exams. For most of the exams, MLO and CC views of both right and left breasts are provided, resulting in 28294 DM images in total. The sample comprises 25901 normal or benign images and 3023 malignant lesions. All images with malignant lesions were histopathologically confirmed, while normal exams were selected if they had at least two years of negative follow-up. From 7196 DM exams, 2883 exams (42%) contained a total of 3023 biopsy-verified malignant lesions. The exact distribution of the dataset is shown in table 2. Furthermore, lesion masks were provided together with the images. Moreover, 1315 exams does not have either left or right breast images of MLO and/or CC views.

Training, validation and test data split was done at exam level to evaluate the generalization of the model developed. Data were randomly split into training (50%), validation (10%) and testing (40%) while making sure exams from each vendor present in each partition proportionally.

Images were energy band normalized Philipsen et al. (2015) and down-scaled to 200 microns after applying a Gaussian filter to homogenize the pixel size across different vendors.

3.2. Candidate Selection Stage

In this study, we have proposed a two-stage breast cancer detection system in which the first stage detect suspicious candidates center location, followed by CNN based approach to reduce FP candidates. Suspicious candidates in the first stage were detected using the previous work of (Karssemeijer, 1999). We refer to this step as the candidate step. Likelihood of a pixel to be part of a mass was computed using local lines and distribution of gradient orientation features. Then, a global threshold was applied to the likelihood image to generate regions that are considered as suspicious. Figure 2) shows sample MLO view DM images of left and right breast. The right breast images are flipped horizontally to put both images in the same space. The red and green points correspond to suspected candidate center locations of mass. While the candidate marked green is a true mass, the others are false predictions. The role of the second stage will be to reduce these false candidates.

Table 3 presents the number of suspicious candidates obtained in stage one for training, validation and test data. 339725, 62926, and 256447 points were found to be suspicious lesion center from the training, validation and test exams respectively, which are outside the lesion mask. 2275 candidates from the training DM were found to be inside the lesion mask, while 1047 and 794 were for validation and test data.

3.3. Patch Preparation

To train our baseline and symmetry CNN, patches centered on the points detected in Section 3.2 were extracted as shown in figure 2. To extract symmetry patches, first the right breast mammography image was flipped horizontally. A simple mirror mapping was applied to locate patch boundary on the contra-lateral image as shown in the figure. Similarly, candidates location of right breast were transformed according to equation (1).

$$(C_x^t, C_y^t) = (R - C_x, C_y) \tag{1}$$

where R is width of the image, and C_x and C_y are candidates x and y location before transformation. C_x^t and C_y^t are transformed coordinates.

Table 1: Previous works on breast mammography lesion, mass and calcification detection, and classification. Task stands for the target task of the study: BI= Breast Imaging Reporting and Data System(BI-RADS), M = Mass detection, L: Lesion, μC: Micro-calcification. In the images column, the total number of images used and the number of test images is displayed in parenthesis. Im. size is the size of image or patch used to train a convolutional neural network. Metric is evaluation method used in the study. The number of stages details whether the experiment is completed in a single stage or multiple stages. Column Info displays the information given to the network: SI: Single input, MV: Multi-view, Sym: Symmetry, Temp: temporal data. Input is presented whether the whole image (WI) or Patch (P) was used as an input. D*: (TPR/image,FPR)=0.87, 0.8, A* = (auc=0.87(μC), 0.9(μc+mass)), B* = (acc=0:929, recall:.934) , acc: Accuracy, sen:Sensitivity,

Ref.	Task	Images	Im. size	Metric	Stages	Info	Input
Carneiro et al. (2015)	BI	680(340)	264x264	auc(0.91)	1	SI	WI
Zhu et al. (2017)	BI	410(CV)	224x224	auc(0.90)	1	SI	P
Arevalo et al. (2016)	L	736(300)	150x150	auc(0.826)	2	SI	P
Huynh et al. (2016)	L	607(cv)	512x512	auc(0.86)	2	SI	P
Carneiro and Bradley	M	410(cv)	264x264	D*	2	SI	P
Geras et al. (2017)	BI	829k(57k)	2600x2000	auc(0.787)	1	MV	WI
Ertosun and Rubin (2015)	M	2500(250)	256x256	acc(85%)	1	SI	P
Mordang et al. (2016)	μC	1606(378)	13x13	sen(0.6914)	2	MV	P
Domingues and Cardoso (2013)	M	116(cv)	32x32	acc(0.86)	2	SI	P
Lévy and Jain (2016)	M	1820(182)	224x224	B*	2	SI	P
Jiao et al. (2016)	M	600(CV)	227x227	acc(97%)	2	SI	P
Kooi and Karssemeijer (2017)	M	73464(18366)	250x250	auc(0.895)	2	Sym and Temp	P
Dhungel et al. (2017)	L	410(CV)	120120	auc(0.80)	1	MV	P
Wang et al. (2016a)	L	1204(204)	1024x1024	A*	1	SI	WI

Table 2: Distribution of DM dataset used

	General Electric	Siemens	Hologic
number of studies	2248	1518	3430
normal images	7771	5842	12288
images with malignant lesions	1292	255	1476

The green box in 2 represents a rectangular patch centered on a positive candidate on MLO view of left breast image, and its corresponding symmetry patch at the same location on the contra-lateral image is displayed in blue. These patches will be used to train a second stage CNN which incorporates symmetry information. In our training dataset, the largest sized mass was 5cm (250 x 250 pixels) and patch size of 6cm (300 x 300 pixels) was selected.

The largest size of a mass in our dataset is 5cm, computed from masses and we selected a patch size of 300x300 pixels (6cm x 6cm). Figure 3 shows sample malignant and normal patches.

3.4. Sampling Strategies

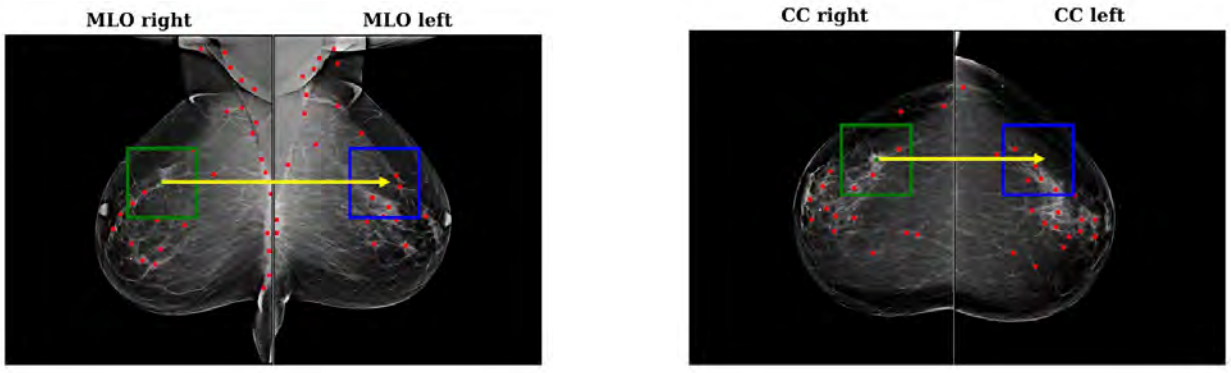
In average, 24 suspicious candidates were detected per image resulting in 659098 candidates of which 3837 (0.6%) are positive candidates. Thus, the pathological observations were remarkably less compared to healthy candidates. This causes a serious problem during training especially if objective function used is not robust to class imbalance. Thus, we under-sampled the negative patches in the training data. All negative candidates in an image are analyzed one by one and considered for training if it satisfies the following conditions:

1. Candidate is at a Euclidean distance of greater than 100 pixels from a lesion (if any).
2. If another candidate within 70 pixels Euclidean distance is not considered. For every candidate in the image, a Euclidean distance is computed to all other candidates and a candidate will be considered if any the candidates which are in distance less than 70 pixels are already in the training patch list.

The 100pixels and 70pixels threshold distances were

Table 3: Distribution of suspected candidates from first stage. The numbers after + indicates candidates from exams without left or right breast images.

Candidates	Training	Validation	Test
Negative	337366+2359	61833 +1093	250293+6154
Positive	2217+58	927+30	727+67



a) b)

Figure 2: An illustrative example showing center location of suspected masses in a sample MLO and CC views of right and left breast mammogram images, and patches used to train symmetry CNN model. The first image is left breast mammogram image while the second is for right breast. Green points represents positive candidates, and red markers are negative. a) MLO view and b) CC view

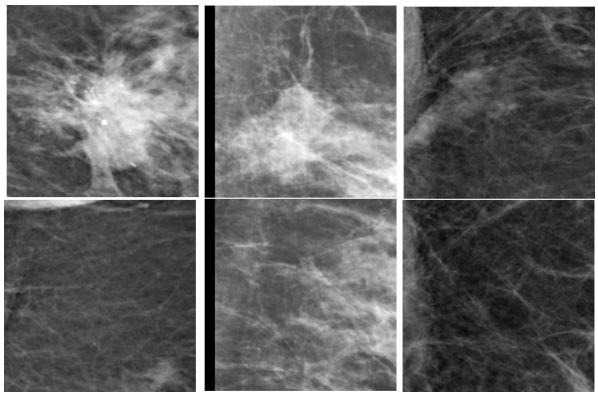


Figure 3: An example of negative (last column) and positive (first two columns) samples mammogram images. Patches in the top row are from the main image and their corresponding symmetry patches are in the bottom row.

chosen empirically. This reduced the number of negative patches to 253476, which is 74.6% of the original amount. However, still, the proportion of positive to negative is 1 to 112 respectively. Thus, further patch balancing is applied during training as explained in Section 3.6.

3.4.1. Preprocessing

The images were already energy band normalized and scaled to same resolution 200microns. After patches are extracted, pixel intensity normalization to be in the range [0,1]. Normalization reduces training time by allowing larger learning rate and makes training less sensitive to weight initialization (Ioffe and Szegedy, 2015).

3.5. Patch Augmentation

Augmentations refer to generating new images from the available images by applying transformation, deformation and adding noise to alleviate data scarcity and overfitting by incorporating more variation of the data at hand. However, the images should be realistic. In a mammogram, the main variations at lesion level mainly scale, rotation, translation and amount of occlusion of tissue. The lesion can be detected at a different stage and this was incorporated by training on a scaled version of input patches, though, later stage lesions might not be simply scaled version of early stage masses Kooi et al. (2017). Moreover, our problem is translation invariant, as lesion center by candidate detector described in section 3.2 can be at any part of the lesion.

Initially, the positive patches are flipped (up-down and left-right) and Gaussian blurred with a standard deviation ranging from 0.2 up to 3. Blurring was applied only to the original patches, not to flipped patches. These generated patches were saved on a disk.

Then, both negative and positive patches (augmented + original) were augmented in real-time. Training was done using Keras Chollet et al. (2015) fit_generator, a keras model method that trains and optimizes a model on a batch by batch data generated by Python generator. Keras has a built-in image data generator class that supports real-time images augmentation. The main function of the generator is to generate a batch of images and their labels according to the setting given at every iteration of training batch. The main downside of this generator is that in case there is large data imbalance, selected batch of images can be from one class which results in overfitting.

Thus, we implemented an image generator which selects an equal number of positive and negative candidates in a batch and applies augmentation to every candidate with a given probability p. In our case we selected $p = 0.5$, chosen empirically. Then, if a candidate was chosen to be augmented, a randomly selected augmentation will be applied to scaling, translation, and rotation. For the purpose of shrinking and translation, originally, patches of size 350 pixels were extracted. Scaling range was set to [0.88, 1.25] inclusive, thus depending on the selected value, the patch will be either up-scaled or down-scaled. In case of translation, the center of the lesion is transformed by a value randomly selected from the range [-25, 25] both in the x and y-direction. The translation in the x and y are selected independently. Rotation is another commonly used augmentation both in computer vision and medical application, and to make it robust, the rotation angle was chosen in the range [-30, 30].

Rotation, translation, and scaling were done in real-time. This saves the need for a large storage memory, however, it slows training time especially when batch size is large. Multiprocessing was used to speedup augmentation time. We have used a batch size of 64 and created 4 processes, in which each process will read 16 patches and apply augmentation. Finally, data from all processes will be combined and fed to the network for training. For symmetry model, the main and symmetry patches are read and augmented in the same way by the same process.

3.6. Network Architecture and Training

In addition to symmetry CNN, a baseline architecture shown in figure 5 was built and trained. This baseline model is used as a reference to evaluate the performance gain of incorporating symmetry information.

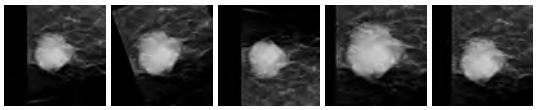


Figure 4: sample augmented images. The first image is original patch. The other images from left to right are generated by rotation, flipping, zooming and translation respectively.

3.6.1. Baseline Architecture

It is a variant of VGG architecture Simonyan and Zisserman (2014) as shown in figure 5 and it consists of feature extraction and classification part. The feature extraction section has a series of seven convolutional layers with $\{16, 32, 32, 64, 64128, 128\}$ neurons and max-pooling layers. Convolution was performed with a stride of (1,1) and valid padding. The classification part is composed of three dense layers of neurons $\{300, 300, 2\}$. Dropout (rate of 0.5) was used after first dense layer. Relu activation was applied to all layers except the last layer, where softmax was applied. The depth of the network and the number of layers were found using a grid search considering the comments of (Kooi and Karssemeijer (2017), Geras et al. (2017)).

Global average pooling (GAP) Lin et al. (2013) was applied after the last convolutional layer while the other layers are followed by 3x3 max pooling. The advantage of GAP over flattening is that it minimizes the chance of overfitting by reducing the number of parameters.

3.6.2. Symmetry Model

The symmetry model has two inputs, a candidate under consideration and a patch on the same location from a contra-lateral image as shown in figure 6. The parallel streams are transfer learned from baseline network in figure 5. Then, the features are extracted from both inputs and concatenated before feeding to the classifier. The classifier part setup up in the same way as the baseline model. The advantage of transferring the learned weights from the baseline architectures is that the network converges in a short time.

Features from the parallel data streams can be fused at any level and to the best of our knowledge, there is no work that investigates this. Most of the parameters in CNN based classification methods comes from the fully connected or dense layers. Therefore, if the parallel data streams are fused at an early stage, the dimension of the input tensor to the fully connected will be high. This results in a dramatic increase in the number of hyperparameters, and a high chance of overfitting. The number of hyperparameter in a dense layer with a given number of neurons is linearly proportional to the dimension of the input tensor. Thus, although we did not perform an experimental investigation, we strongly believe that it is advisable to fuse the features at a later stage which gives a reasonable number of hyperparameter taking the

amount of training data.

Multi-input CNN can be trained using the single stream and multiple data stream architectures. In the study by Kooi and Karssemeijer (2017), it is stated that a multi-stream network outperformed single input network and in this work, we have proposed the network presented in figure 6 without sharing weights of the parallel data stream. The output of the max-pooling after the last convolutional layer is three dimensional (3x3x128). It is flattened before feeding to the classifier.

One of the most common problems in machine learning is missing data. It is an old problem and there have been a lot of imputation techniques to handle this in statistical machine learning, such as k-nearest neighbor Batista and Monard (2003), mean, mode and predictive replacement Poulos and Valle (2016), zero imputation and forward filling in the context of recurrent neural network(Lipton et al., 2016). In zero imputation, missing data are replaced by zero and missing values are set to a previous value in case of forwarding filling. For imputed data, missing data perturbation can improve generalization of the model by regularizing the model Poulos and Valle (2016), and zero imputation outperformed forward filling in Lipton et al. (2016).

In our dataset, there were exams with no right or left breast DM images, and we considered them as missing data. For the image from these exams, zero matrices (image) was used as a contra-lateral image. Quantitative evaluation of symmetry model on exams with missing data is presented in Section 4.

3.6.3. Training

Weights are initialized using Glorot weight initialization Glorot and Bengio (2010) and optimized using Stochastic Gradient Descent (SGD) with time-based learning rate scheduler with an initial learning rate (ILR) of $1e^{-2}$ for baseline network and $1e^{-3}$ for symmetry model, with decay rate of ILR/200, and momentum value of 0.9. Mini-batch size of 64 was used.

During training, categorical cross entropy objective function was optimized. For a mini-batch size of M patches the computed cross entropy loss is given by equation (2).

$$loss = -\sum_{M}\sum_{i} y_i log(\hat{y}_i) \tag{2}$$

where M is batch size, y_i ground truth class and $\hat{y}_i$ is predicted probability.

As explained in Section 3.3, in our dataset, the number of positive patches are much smaller than negative patches. Thus, if a batch of images is randomly selected the images might be only from negative class and this results in overfitting. Buda et al. (2017) has stated that the effect of class imbalance is detrimental for classifier performance and oversampling performs better. However, if the imbalance cannot be completely removed, under-sampling outperforms Buda et al. (2017). In our case, as the imbalance is 1:150, oversampling will not work. Instead, oversampling and under-sampling were applied.

Moreover, additional strategies were applied. First, an equal number of positive and negative were fed to the network at every iteration using a generator. The number of iteration per epoch was determined by the positive samples, the underestimated class. Every positive sample was seen multiple times per epoch similar to the work of Kooi et al. (2017), however, it did not explicitly specify the number of times. We evaluated our baseline model for a different number of repetition

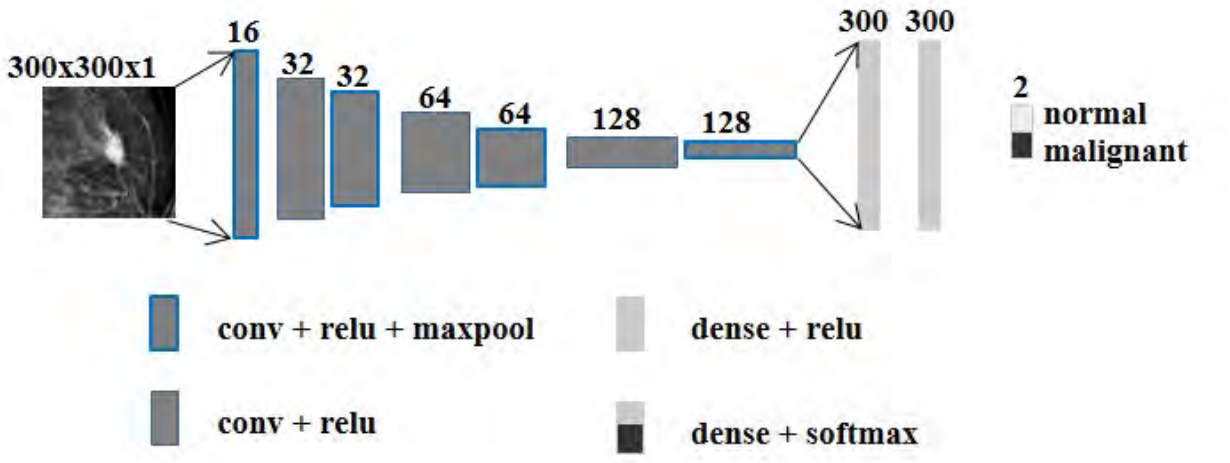


Figure 5: Illustration of baseline CNN architecture; it is scaled down version of VGG16 Simonyan and Zisserman (2014). The feature extraction part is composed of a series of convolutional and max-pooling layer. Dense layer with relu activation were used as classifier. Each convolution has 3x3 kernel. The output of the classifier is binary, normal or malignant patch.

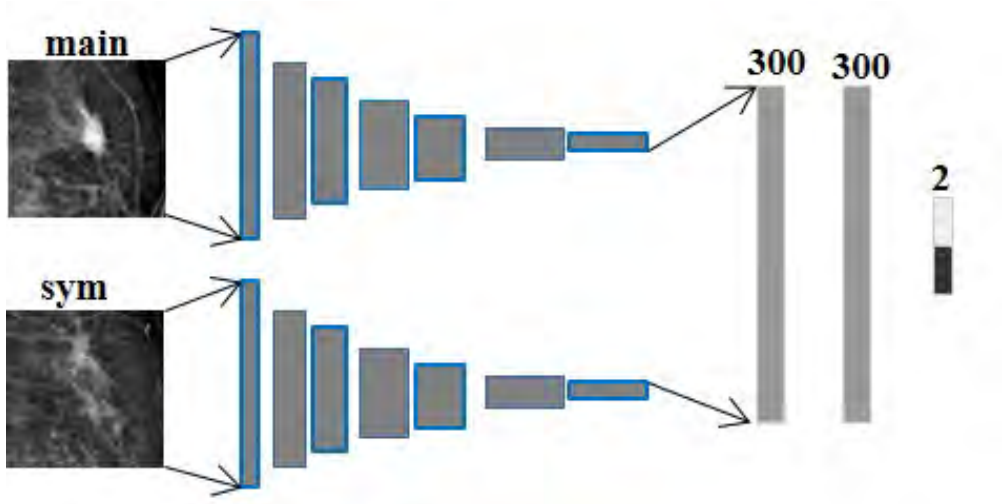


Figure 6: Symmetry model architecture. It has two inputs of gray scale 300 X 300 pixels size patches with their own data stream for feature extraction. Then, the features are concatenated and fed to a neural network classifier which contains a series of dense layers. Input main is the a patch from main image and input sym is a patch from contra-lateral image

and we found that repeating twice succeeds in terms of generalization and training time. This approach has the following advantages; first, it improves generalization of the model as the weight update is based on the prediction loss of an equal number of samples from both classes. Second, reduces the time per epoch as the number of iteration is smaller. Furthermore, training with the same image many times will result in overfitting as the network learn the noise in the underestimated class. This is the same problem as a class or sample weighting when the class imbalance is large.

Secondly, when a given sample is seen repeatedly in a given epoch and at different epochs, it is likely to be its augmented version as there is runtime augmentation as explained in Section 3.5.

Moreover, we have created a custom callback function that computes AUC at the end of every epoch on validation and sampled training patches. As the training set is large to minimize the computation time, 30% of the training set was sampled and used as a representative for the whole training data. Then, AUC was computed at every two epochs interval and the best model was selected based on AUC on validation patches.

Finally, early stopping was applied. During the progress of training, training loss continues improving, however, validation loss starts to drop after some epochs due to overfitting. Early stopping attempts can be considered a type of regularization method in that it can stop the network from overfitting. We monitored AUC for early stopping and patience was set to 20 epochs.

3.7. Performance Evaluation Metrics

One of the most frequently used performance metric for classification problems is AUC, especially when there is skewed sample distribution like in our problem. Moreover, detection and localization of one or more targets is a common task in diagnostic image analysis and other computer vision problems. In such cases, FROC is also used as a diagnostic performance evaluation method (Bandos et al., 2009). Receiver Operating characteristic (ROC) curve gives evaluation at a candidate level. In contrast to ROC, FROC gives information to asses the performance at an image or exam level and takes into consideration the number of targets. In this work, candidate and lesion level performance comparison were done using ROC and AUC, whilst image and exam based evaluation were performed with the help of FROC and CPM. CPM score was introduced by (Niemeijer et al. (2011), (Setio et al., 2017)) and it is computed as an average sensitivity from the FROC at seven False Positive Rate (FPR): 1/8, 1/4,1/2,1,2,4, and 8.

Image-based FROC is computed per image and used for image level performance evaluation, while exam basedFROC is computed per exam, considering all views of left and right breast images.

FROC curve is plotted True Positive Rate (TPR) against FPR per image or FPR per exam and the points on the FROC curve are computed as follows:

1. For all candidates in test dataset, the prediction probability, and ground truth label was recorded.
2. All unique probabilities were sorted and marked as a threshold (T) to compute TPR and their respective FPR
3. Then, TPR and FPR were computed at every threshold, T. FPR is computed only from normal exams as we can guarantee there is no lesion.
    - For a image based FROC, for every T, FPR

was computed as average number of FPR per image, where FPR per image is a ratio between the number of negative candidates wrongly categorized as positive (FP) and the total number of actual negative candidates. TPR is determined as an average TPR per image, where TPR per image is the proportion of True Positive (TP) that are correctly predicted.

- For case based FROC, FPR is computed in a similar way as image based FROC, but per exam. TPR is set to 1 if at least one of the lesions in the exam are correctly classified and else zero.

Competition performance metric is an average sensitivity computed at 8 predefined different FPR per image or FPR per exam depending on where the performance is being evaluated. Different CAD systems use different FPR threshold. An interesting point about CPM is that low and high FP are considered. This determines if the model can also significantly identify masses at a small FP predictions.

Most machine learning algorithms such as CNN based approaches suffer from randomness and selecting the best model is a challenging task (Brownlee, 2015). In this study, 95% confidence interval of AUC and FROC CPM was computed using bootstrapping as described in (Efron and Tibshirani, 1994), with 1000 bootstraps, to compare the performance candidate selection stage, baseline architecture and symmetry model. Moreover, a p-value of our models was computed to measure the significance of the difference in performance at different levels. Significance test and p-values were computed on our test data using 1000 iterations employing sampling with replacement, each iteration taking 70% of the whole testing data. To figure the confidence interval and p-value at exam level, sampling was done at an exam level. similarly, for candidate level analysis, sampling was done at a candidate level. In many statistical analysis p-value smaller than 0.05 was considered as significant.

4. Results and Statistical Analysis

We first presented an empirical analysis of our CNN based models and candidate selection, then, qualitative results of best-performing symmetry CNN model is presented. All the experiments were conducted in KerasChollet et al. (2015), and all the results presented here are on a separately held 40% of our dataset. Quantitative evaluation was done at candidate, lesion, image and exam level.

4.1. Candidate and Lesion Level

In most of the images, during candidate detection, more than one suspicious candidates were selected inside one lesion area. Candidate level evaluation refers to the performance of the models for every candidate detected in the first stage. For lesion level evaluation, predicted probabilities of all candidates detected within the region a given lesion were grouped and the maximum malignancy prediction probability was chosen as malignancy probability of the lesion.

Figure 7a shows candidate level ROC and their respective AUC values for the three models. Moreover, 95% confidence interval of the all the models is detailed in Table 4. At candidate level, AUC value of 0.896 with confidence interval of [0.879 , 0.913] was obtained by the model used for candidate selection. The baseline architecture that processes a single ROI image yield AUC value of 0.929 with confidence interval [0.916, 0.942], significantly higher than the candidate selection stage performance (p = 0.004). Incorporating symmetry information undoubtedly improved the AUC to 0.933 with [0.919 , 0.947] 95% confidence interval, although it was not significant (p = 0.111) in comparison with baseline architecture.

Figure 7b presents ROC curve comparison of symmetry model evaluated on all test data and on images with missing contra-lateral part. The evaluation on missing data is based on 6152 negative and 67 positive candidates (Table 3) and AUC value of 0.866 was obtained with 95% significance interval of [0.788, 0.930].

From figure 8, AUC values of 0.943, 0.937, and 0.935 was scored by the first stage, symmetry and baseline model respectively when performance is evaluated at lesion level and their confidence interval is also shown in the second column of Table 4. At lesion level, the performance of the first stage was higher compared to symmetry and baseline model (p = 0.216 and p = 0.082 respectively). In contrast to candidate level evaluation, the AUC value candidate selection stage was found higher than the CNN models. As the maximum predicted probability from all the candidates that lie in a given lesion was assigned as a malignancy probability, the algorithm used in a candidate level has managed to get slightly higher sensitivity, although the difference was not significant.

4.1.1. Predicted Probabilities Distribution

To understand the distribution of the predicted probabilities of suspected candidates, the mean(μ) and standard deviation (σ) of predicted probabilities for the negative and positive candidates are displayed in figure 9 for the three models. For candidate selection stage, the mean of predicted probabilities of negative and positive candidates was found 0.6402 and 0.366 respectively. The average predicted probability obtained for positive candidates using baseline and symmetry architectures is slightly higher than the first stage, 0.724 and 0.673 respectively. For negative candidates, the CNN models have a remarkably smaller mean prediction, however, the variance was a bit higher in case of CNN models. From the bar graph, it is clearly evident that symmetry

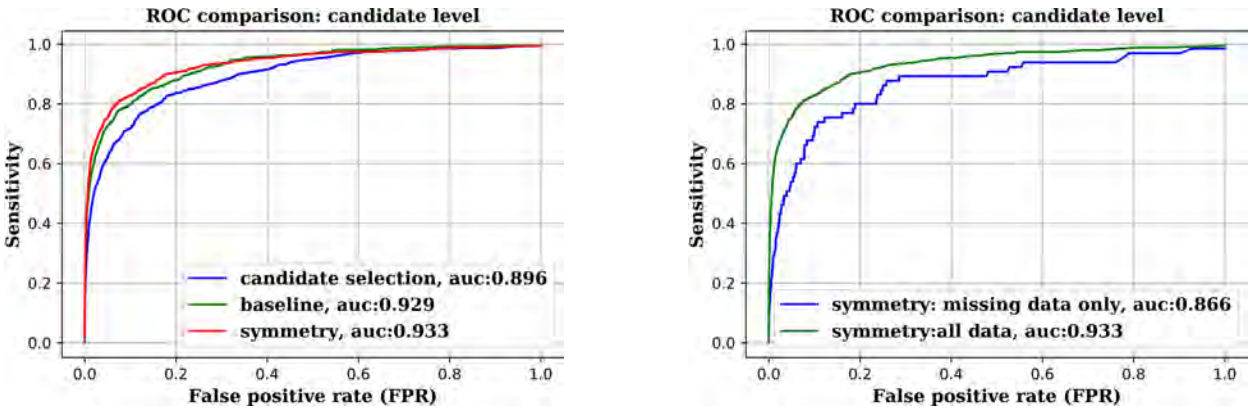


Figure 7: Candidate level ROC curve comparing discriminative performance of the three models. a) Compares ROC curve and AUC value of candidate selection stage, baseline CNN and symmetry CNN. b) Compares ROC and AUC evaluation the best performing symmetry model on candidates with missing contra-lateral image and overall performance.

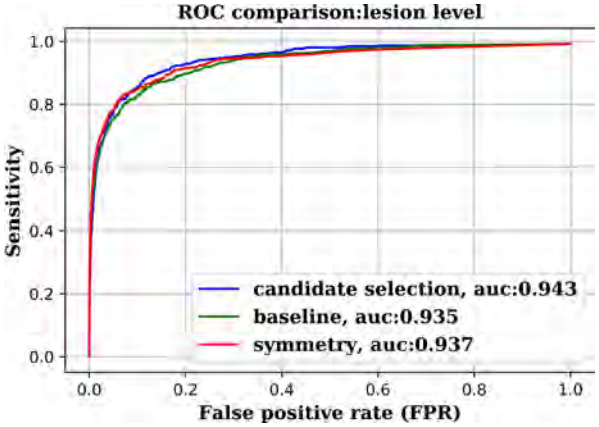


Figure 8: Lesion level ROC curve comparing discriminative power of candidate selection, baseline and symmetry models.

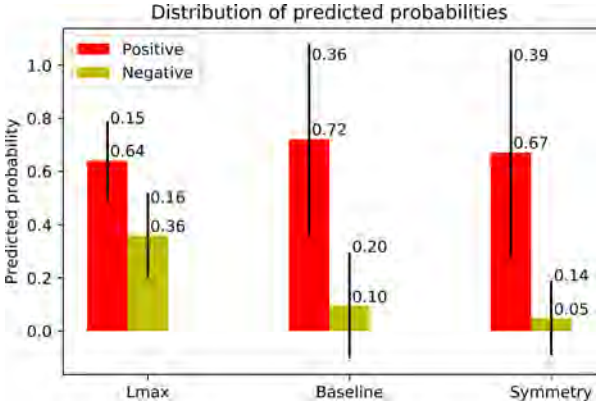


Figure 9: Distribution of predicted probabilities of suspected candidates. The numbers displayed are mean and standard deviation values, and the black thin bars show standard deviation of the values around the mean.

model was better in reducing FP candidates as the mean malignancy prediction probability was found smaller in comparison with the other models.

4.2. Image and Exam Level

Figure 10a and 10b present image and exam based FROC along with CPM values comparison of the three models and 95% confidence interval of CPM values computed from FROC using 1000 bootstraps are shown in Table 4. In our test set, symmetry model showed a reasonably better performance (p = 0.001) compared to the single ROI baseline architecture in both image and exam level statistical analysis. Competition performance metric value of 0.716, 0.718, and 0.744 with 95% confidence interval of [0.682 , 0.750], [0.679 , 0.756] , and [0.723 , 0.794] was obtained for candidate selection, baseline and symmetry model respectively when evaluation was done at an image.

Moreover, during exam level evaluation sensitivity of our model that incorporates symmetry information was found higher than the other model throughout the whole range of FPR, resulting in CPM value of 0.733 compared to 0.682 and 0.702 for candidate selection and baseline model, respectively.

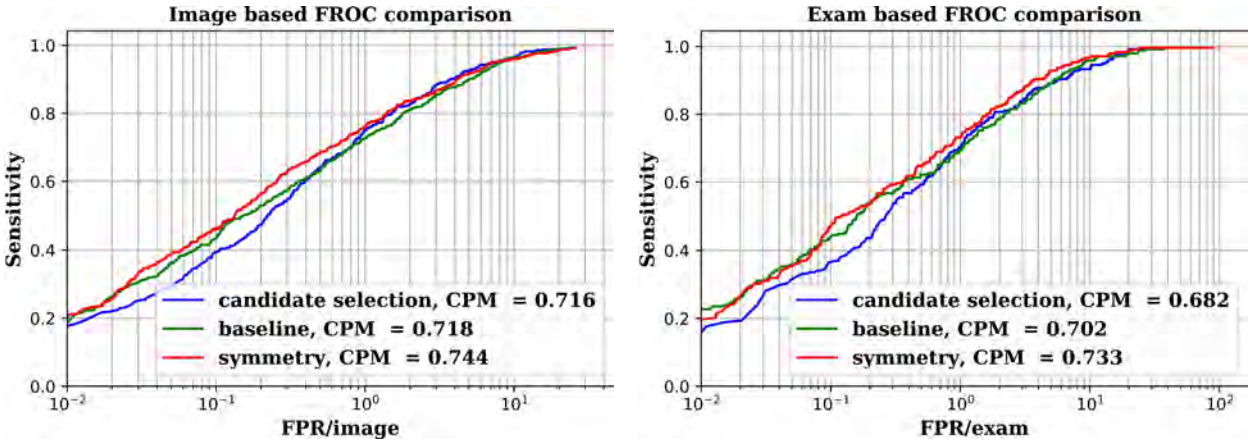
From figure 10a and figure 10b, for FPR less than 0.3, CNN based models have outperformed in comparison with candidate selection stage model. For FPR greater than 0.5 candidate selection stage and baseline model have similar performance while symmetry model has continued to perform better.

5. Discussion

In our dataset, there is a remarkable amount of imbalance between positive and negative candidates (1 positive to 150 negative candidates). To mitigate this, we

Table 4: 95% Confidence interval. For candidate and lesion level evaluation, the values indicated corresponds to 95% confidence interval of AUC and CPM was used for image and exam level evaluation.

Model	Candidate	Lesion	Image	Exam
Candidate selection	[0.879, 0.913]	**[0.925, 0.957]**	[0.682, 0.750]	[0.671, 0.746]
Baseline CNN	[0.916, 0.942]	[0.917, 0.949]	[0.679, 0.756]	[0.713, 0.772]
Symmetry CNN	**[0.919, 0.947]**	[0.920, 0.954]	**[0.723, 0.794]**	**[0.721, 0.823]**



a) b)

Figure 10: FROC comparison of candidate selection, baseline and symmetry models; a) Image based FROC comparison. b) Exam based FROC comparison. The y axis represents sensitivity in linear scale and x axis FPR per image or exam in logarithmic scale. Lesion level prediction probabilities were used to compute FPR and TPR.

have experimented with different sampling strategies on the overestimated class. In the whole training set, there were 2309 positive candidates and sampling was applied only to negative class. The number of negative samples after the different strategies and AUC value of best performing baseline models on the validation set is presented in Table 5. When negative candidates are taken from normal exams as in the work of Kooi and Karssemeijer (2017), 322062 candidates were collected after sampling. Then, the baseline model in figure 5 is trained and the maximum value of AUC obtained was 0.917. Sampling strategy described in Section 3.4 was applied to the other methods in the table. When the threshold is 70, the number of candidates was reduced to 253476 and 264211 for a threshold value of 50.

When the threshold value is 70 AUC value of 0.926 was obtained compared to 0.922 and 0.917 for a distance threshold value of 50 and sampling from normal cases, respectively. A model trained on with negative candidates from positive and negative exams outperformed as more variety of candidates capturing the real distribution of the class were incorporated. Another advantage is that faster training time due to a smaller number of data while maintaining the variance of the data. These sampling strategies will minimize redundant information. It should be noted that increasing the distance threshold value does not guarantee an increase in performance in other problems. This depends on the variety of patches and the distance of candidates from the first stage candidate selection.

Undersampling described above cannot completely solve data imbalance we have. As in the work of Farid and Rahman (2013), sample or class weighting increases classification accuracy, however, in our case, the result got worse. The under-sampled class was given large weight while smaller weight was associated with the overestimated class or samples. We have realized that sample and class weighting works when the degree of imbalance is not very large, which is not the case in our dataset. The loss gradient is a weighted average over the batch and that if all of the batches contain at least one positive sample, the gradients will bias to overfit to the positive class due to the large weight. This drawback might be minimized by increasing batch size.

We have solved this by applying real-time geometric transformation during training from a continuous range of values as described in section 3.5 and by feeding augmented version of positive samples twice per epoch. Table 6 shows convergence epoch number for a different number of repetitions (R) on the positive class to reach AUC value of 0.92 on a validation set.

The larger the number of repetition, the smaller number of epochs to converge. However, it should be noted that the training time per epoch will be larger and there is a high chance of overfitting to positive candidates. when the number of repetition is 8, the network gave AUC value of 0.92 after 2 epochs, and when the positive case is seen only once convergence is reached after

Table 5: AUC performance of different data sampling strategies using one view model. Normal exams refer to mammography exams without lesion and distance specified are Euclidean distance.

Sampling strategy	AUC	Number of negative candidates
Only from normal exams	0.917	322062
Distance threshold, 50	0.922	264211
Distance threshold, 70	0.926	253476

26 epochs. So, to minimize the chance of overfitting and reach convergence at a reasonable time, for the rest of our experiments, R=2 was chosen.

Table 6: Number of epochs needed to reach AUC value of 0.92 for different number of repetition. Number of repetition is the number of times that a positive candidate is presented in a given epoch.

Number of repetition (R)	Number of epochs
1	26
2	6
4	4
8	2

In this study, we have trained the baseline architecture from scratch. As the primitive patterns that are useful for malignancy detection in one view model will be important even when symmetry image is added, the weights of the parallel data streams in the symmetry model were initialized by the weights of pretrained baseline architecture. This has helped us to train on large-scale data in a short time and get better performance within a small number of epochs. One of the drawbacks of using multi-stage approach is candidates missed in the early stages cannot be recovered in the later stage and the performance of the later stage will be lower as its inputs are those picked by the previous stage. In our case, there were positive candidates that were missed in the first stage and this implies even at FPR of 1.0, sensitivity will be slightly less than 1,0. For example, in figure 8, sensitivity of 0.994 was achieved at FPR value of 1.0 due to missed lesions in the candidate selection stage. Therefore, having a candidate detector with high sensitivity is crucial for the overall performance of the CAD system.

To investigate what the network is learning and missing, we have visualized some of the misclassified patches as shown in figure 11.

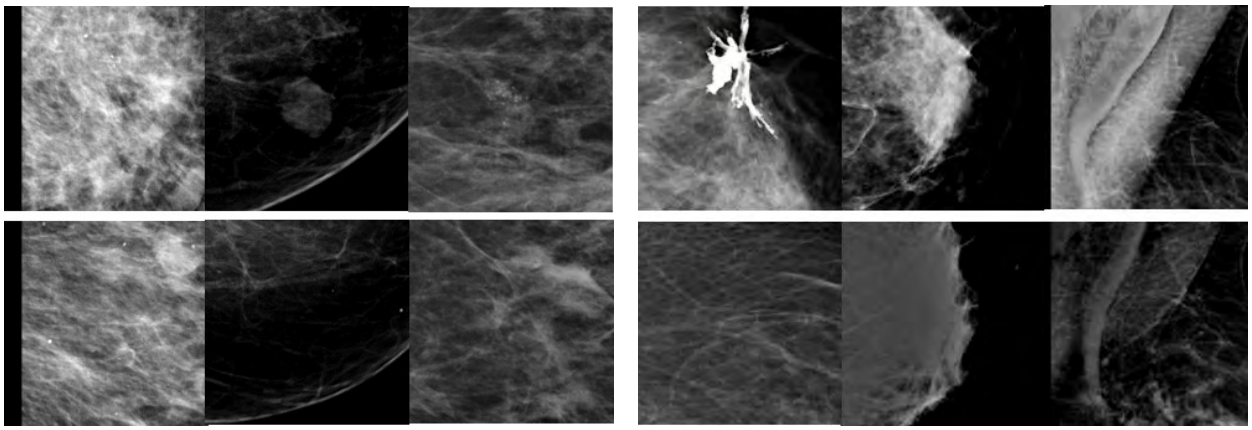
Figure 11a shows sample patches with a biopsy-proven lesion, but the network predicted to have a lesion with probability less than 0.1. Most of the misclassified positive candidates include large lesion, a lesion that looks benign and/or patches with micro-calcification. Larger lesions were underestimated during training. Benign looking malignant candidates could be better discriminated by including temporal data.

Negative patches that were predicted to be malignant lesion with a probability of at least 0.9 are displayed in figure 11b. Most of these patches comprise benign abnormalities such as cyst and fibroadenoma, normal structures like fat necrosis, breast nipple, and pectoral muscle, and patches with artifacts. In some of the exams large part of the pectoral muscle is visible, but not in the other breast. Thus, when the main patch is inside pectoral muscle but not the symmetry, the network is predicting as a lesion, because of asymmetrical difference. Moreover, the left and right DM images are not also aligned and this affects the performance of symmetry model. Alignment and pectoral muscle problem can be solved by proper scanning when the DM are obtained. The benign FP candidates can be filtered by including third class, benign in the framework.

In this work, we have extensively employed data augmentation using geometric transformation. We use rotation, translation, flipping, and scaling. These augmentations are justified as the mass classification is invariant to these transformations. To avoid artifacts outside the boundary during rotation and for real-time translation, 350x350 pixels sized patch was extracted and cropped from the center to 300x300 pixels when it is fed to the network. One of the drawbacks of restricted augmentation is that the network might learn augmentation even if the model selection is based on validation set as the weight update is based on the loss of training batch.

We have found that perturbation of augmentation type and parameters through real-time data generation improves the generalization of the model and thus, the performance of the classifier. The more flexible is the augmentation type and the more generalized will be the model. Symmetry model trained without patch augmentation described in Section 3.5 yielded AUC value of 0.91 on a test set, in comparison to 0.933 when augmentation was applied. Real-time augmentation slows the training time, and we managed it using multiprocessing to speed up augmentation time. Moreover, images are copied from the server to GPU machine multiprocessing.

We have found that incorporating symmetry information helps in learning distinctive features when there is a low-intensity contrast between mass and the background as shown in figure 12a. Baseline architecture (without symmetry information) failed to discriminate these patches 12. the malignancy probability was found to be below 0.2, however, integrating symmetrical information increased malignancy prediction to a value greater than 0.7. Moreover, the images in figure 12b



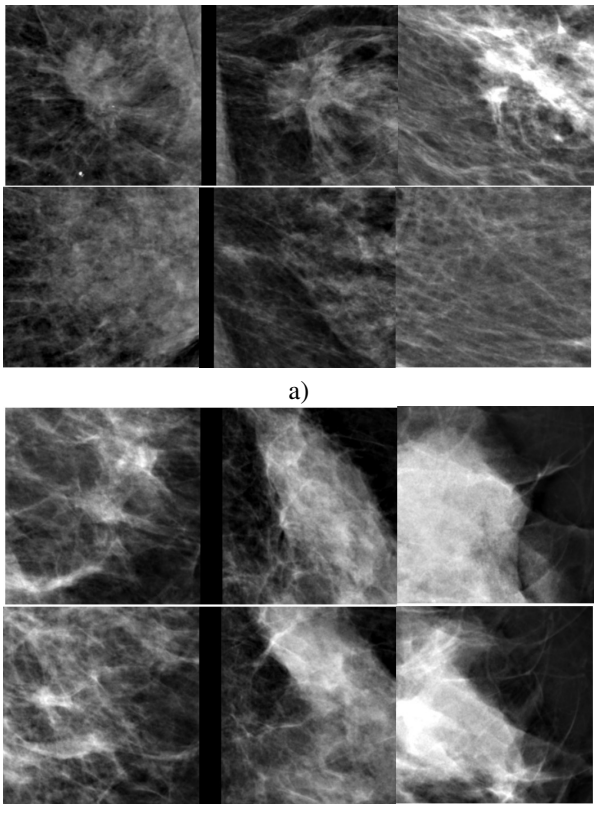
a) b)

Figure 11: An illustrative example of top misclassified patches by symmetry model. The top and bottom row images correspond to primary and contra-lateral patches, respectively a) patches with a mass that were predicted with malignancy probability of less than 0.1. b) negative patches that were predicted as positive

were predicted as malignant masses by the baseline model with probability greater than 0.9. In mammogram images, it is less likely to find a lesion at the same location, and as the primary and symmetry images look similar, symmetry model has manages to discriminate them as normal patches.

As shown in Table 2 in the previous work by Kooi et al. (2017), which was focused on soft tissue lesion classification, it is reported that AUC value of 0.895 was obtained by incorporating symmetry and temporal context information. Although a direct comparison might not be convenient with state of art methods in Table 1 as a different dataset was used in this study, our approach has the highest AUC value, which is 0.933 at candidate level and 0.937 at lesion level.

One of the main limitations of this work is that only breast mammogram masses are studied and detecting calcification will be an added value. Secondly, as described above some benign abnormalities were difficult for the network to differentiate from malignant candidates. We suggest that separating the benign candidate and training with three classes could improve the detection performance. As studied in Kooi et al. (2017), integrating the different views and more time points could also improve the performance of the model.



a)

b)

Figure 12: Sample patches with an improved prediction using symmetry model: a) positive patches that were misclassified by baseline architecture and correctly classified by symmetry model. b) negative patches that were miss classified by baseline architecture and correctly classified by symmetry model. The top and bottom row images are primary and contra-lateral pairs

6. Conclusion

In this work, we proposed a deep learning approach that integrates symmetrical information to improve breast mass detection from mammogram images. A previous work by Karssemeijer (1999) was used to detect suspicious candidates. The FP were reduced by learning symmetrical differences between primary and contra-lateral patches. Due to symmetrical nature of the breast, for every candidate on the primary image,

a mirror point was considered in the contra-lateral image to train the symmetry model. We have used AUC to measure the performance at candidate and lesion level, whilst CPM were computed for image and exam level evaluation. We have found that our proposed approach reduces FP predictions compared to baseline architecture. AUC value 0.933 (p = 0.111) with 95% confidence interval of [0.919,0.947] was obtained at candidate level and 0.733 (p = 0.001) CPM with 95% confidence interval of [0.721, 0.823] was achieved with our symmetry model. Although a private dataset was used, with our large-scale dataset, we have obtained a better performance than the state of art in terms of AUC and our proposed approach has a potential to be used in breast screening program.

Although an acceptable performance was achieved, training with a dataset which includes more time points could possibly improve reliability and detection accuracy.

7. Acknowledgments

First of all, I would like to thank my supervisors Dr. Jonas Teuwen and Dr. Albert Gubern-Mérida for their guidance, great support, constructive suggestions and kind advice throughout my master thesis.

To all Medical Imaging and Applications(MAIA) community and professors, I would like to express my deepest gratitude for providing a charming educational environment, and constant support with the logistics. I would like to thank all my course professor for your instructive lectures, advice, and support throughout the program.

I would like to thank also for my classmates for their companionship and for creating a friendly environment throughout these years. All the drinks, dinner, trips and classes up to 6 pm will be always remembered. Best of luck for all of you for your future endeavor.

Finally, I would like to thank my parents for their unconditional love, and encouragement to accomplish this master.

References

Anttinen, I., Pamilo, M., Soiva, M., Roiha, M., 1993. Double reading of mammography screening films-one radiologist or two? Clinical Radiology 48, 414–421. doi:10.1016/S0009-9260(05)82949-6.

Arevalo, J., González, F.A., Ramos-Pollán, R., Oliveira, J.L., Lopez, M.A.G., 2016. Representation learning for mammography mass lesion classification with convolutional neural networks. Computer methods and programs in biomedicine 127, 248–257. doi:https://doi.org/10.1016/j.cmpb.2015.12.014.

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 doi:https://arxiv.org/abs/1409.0473v7.

Bandos, A.I., Rockette, H.E., Song, T., Gur, D., 2009. Area under the free-response roc curve (froc) and a related summary index. Biometrics 65, 247–256. doi:10.1111/j.1541-0420.2008.01049.x.

Batista, G.E., Monard, M.C., 2003. An analysis of four missing data treatment methods for supervised learning. Applied artificial intelligence 17, 519–533. doi:10.1080/713827181.

Bick, U., 2014. Mammography: How to interpret microcalcifications, in: Diseases of the Abdomen and Pelvis 2014–2017. Springer, pp. 313–318. doi:10.1007/978-88-470-5659-6_40.

Brownlee, J., 2015. Machine learning algorithms. Machine Learning Mastery .

Buda, M., Maki, A., Mazurowski, M.A., 2017. A systematic study of the class imbalance problem in convolutional neural networks. arXiv preprint arXiv:1710.05381 .

Carneiro, G., Nascimento, J., Bradley, A.P., 2015. Unregistered multiview mammogram analysis with pre-trained deep learning models, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 652–660. doi:https://doi.org/10.1007/978-3-319-24574-4_78.

Carneiro, N.D.G., Bradley, A.P., . Automated mass detection from mammograms using deep learning and random forest .

Chollet, F., et al., 2015. Keras. `https://keras.io`.

Dhungel, N., Carneiro, G., Bradley, A.P., 2017. Fully automated classification of mammograms using deep residual neural networks, in: Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on, IEEE. pp. 310–314. doi:10.1109/isbi.2017.7950526.

Domingues, I., Cardoso, J., 2013. Mass detection on mammogram images: a first assessment of deep learning techniques, in: 19th Portuguese Conference on Pattern Recognition (RECPAD).

Efron, B., Tibshirani, R.J., 1994. An introduction to the bootstrap. CRC press. doi:10.1007/978-1-4842-2382-6_2.

Ertosun, M.G., Rubin, D.L., 2015. Probabilistic visual search for masses within mammography images using deep learning, in: Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on, IEEE. pp. 1310–1315.

Farid, D.M., Rahman, C.M., 2013. Assigning weights to training instances increases classification accuracy. International Journal of Data Mining & Knowledge Management Process 3, 13. doi:https://doi.org/10.5121/ijdkp.2013.3102.

Feig, S.A., 2002. Effect of service screening mammography on population mortality from breast carcinoma. Cancer 95, 451–457. doi:https://doi.org/10.1002/cncr.10764.

Geras, K.J., Wolfson, S., Shen, Y., Kim, S., Moy, L., Cho, K., 2017. High-resolution breast cancer screening with multi-view deep convolutional neural networks. arXiv preprint arXiv:1703.07047 doi:https://arxiv.org/abs/1703.07047v2.

Giger, M.L., Karssemeijer, N., Schnabel, J.A., 2013. Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer. Annual review of biomedical engineering 15, 327–357. doi:10.1146/annurev-bioeng-071812-152416.

Gilbert, F.J., Tucker, L., Young, K.C., 2016. Digital breast tomosynthesis (dbt): a review of the evidence for use as a screening tool. Clinical radiology 71, 141–150. doi:10.1016/j.crad.2015.11.008.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp. 249–256.

Group, S.O.S.S.E., et al., 2006. Reduction in breast cancer mortality from organized service screening with mammography: 1. further confirmation with extended data. Cancer Epidemiology Biomarkers & Prevention 15, 45. doi:https://doi.org/10.1158/1055-9965.epi-05-0349.

Huynh, B.Q., Li, H., Giger, M.L., 2016. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. Journal of Medical Imaging 3, 034501.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 .

Iyyer, M., Manjunatha, V., Boyd-Graber, J., Daumé III, H., 2015. Deep unordered composition rivals syntactic methods for text classification, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1:

Long Papers), pp. 1681–1691. doi:https://doi.org/10.3115/v1/p15-1162.

Jiao, Z., Gao, X., Wang, Y., Li, J., 2016. A deep feature based framework for breast masses classification. Neurocomputing 197, 221–231. doi:https://doi.org/10.1016/j.neucom.2016.02.060.

Karssemeijer, N., 1999. Local orientation distribution as a function of spatial scale for detection of masses in mammograms, in: Biennial International Conference on Information Processing in Medical Imaging, Springer. pp. 280–293. doi:10.1007/3-540-48714-x_21.

Kooi, T., Karssemeijer, N., 2017. Classifying symmetrical differences and temporal change in mammography using deep neural networks. arXiv preprint arXiv:1703.07715 doi:doi.org/10.1117/1.jmi.4.4.044501.

Kooi, T., Litjens, G., van Ginneken, B., Gubern-Mérida, A., Sánchez, C.I., Mann, R., den Heeten, A., Karssemeijer, N., 2017. Large scale deep learning for computer aided detection of mammographic lesions. Medical image analysis 35, 303–312. doi:10.1016/j.media.2016.07.007.

Lévy, D., Jain, A., 2016. Breast mass classification from mammograms using deep convolutional neural networks. arXiv preprint arXiv:1612.00542 .

Lin, M., Chen, Q., Yan, S., 2013. Network in network. arXiv preprint arXiv:1312.4400 doi:10.1002/0471650129.dob0482.

Lipton, Z.C., Kale, D.C., Wetzel, R., 2016. Modeling missing data in clinical time series with rnns. Machine Learning for Healthcare .

Lu, Z., Carneiro, G., Dhungel, N., Bradley, A.P., 2016. Automated detection of individual micro-calcifications from mammograms using a multi-stage cascade approach. arXiv preprint arXiv:1610.02251 .

Martin, A.M., Weber, B.L., 2000. Genetic and hormonal risk factors in breast cancer. Journal of the National Cancer Institute 92, 1126–1135. doi:0.1093/jnci/92.14.1126.

de Moor, T., Rodriguez-Ruiz, A., Mann, R., Teuwen, J., 2018. Automated soft tissue lesion detection and segmentation in digital mammography using a u-net deep learning network. arXiv preprint arXiv:1802.06865 doi:https://arxiv.org/abs/1802.06865v2.

Mordang, J.J., Janssen, T., Bria, A., Kooi, T., Gubern-Mérida, A., Karssemeijer, N., 2016. Automatic microcalcification detection in multi-vendor mammography using convolutional neural networks, in: International Workshop on Digital Mammography, Springer. pp. 35–42.

Niemeijer, M., Loog, M., Abramoff, M.D., Viergever, M.A., Prokop, M., van Ginneken, B., 2011. On combining computer-aided detection systems. IEEE Transactions on Medical Imaging 30, 215–223. doi:10.1109/TMI.2010.2072789.

Oeffinger, K.C., Fontham, E.T., Etzioni, R., Herzig, A., Michaelson, J.S., Shih, Y.C.T., Walter, L.C., Church, T.R., Flowers, C.R., LaMonte, S.J., et al., 2015. Breast cancer screening for women at average risk: 2015 guideline update from the american cancer society. Jama 314, 1599–1614. doi:10.1001/jama.2015.12783.

Oliver, A., Freixenet, J., Marti, J., Perez, E., Pont, J., Denton, E.R., Zwiggelaar, R., 2010. A review of automatic mass detection and segmentation in mammographic images. Medical image analysis 14, 87–110. doi:10.1016/j.media.2009.12.005.

Organization, W.H., et al., 2006. Guidelines for the early detection and screening of breast cancer doi:http://www.who.int/iris/handle/10665/119811.

Philipsen, R.H., Maduskar, P., Hogeweg, L., Melendez, J., Sánchez, C.I., van Ginneken, B., 2015. Localized energy-based normalization of medical images: application to chest radiography. IEEE transactions on medical imaging 34, 1965–1975. doi:10.1109/tmi.2015.2418031.

Poulos, J., Valle, R., 2016. Missing data imputation for supervised learning. arXiv preprint arXiv:1610.09075 doi:https://doi.org/10.1080/08839514.2018.1448143.

Rakhlin, A., Shvets, A., Iglovikov, V., Kalinin, A.A., 2018. Deep convolutional neural networks for breast cancer histology image analysis. arXiv preprint arXiv:1802.00752 doi:10.1101/259911.

Setio, A.A.A., Traverso, A., De Bel, T., Berens, M.S., van den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M.E., Geurts, B., et al., 2017. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. Medical image analysis 42, 1–13. doi:https://arxiv.org/abs/1612.08012.

Siegel, R.L., Miller, K.D., Jemal, A., 2018. Cancer statistics, 2018. CA: a cancer journal for clinicians 68, 7–30. doi:https://doi.org/10.3322/caac.21442.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 doi:10.1109/CVPR.2016.182.

Sree, S.V., Ng, E.Y.K., Acharya, R.U., Faust, O., 2011. Breast imaging: A survey. World journal of clinical oncology 2, 171. doi:10.5306/wjco.v2.i4.171.

Stefan, D.C., 2015. Cancer care in africa: An overview of resources. Journal of global oncology 1, 30–36. doi:https://doi.org/10.1200/jgo.2015.000406.

Tang, J., Rangayyan, R.M., Xu, J., El Naqa, I., Yang, Y., 2009. Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. IEEE Transactions on Information Technology in Biomedicine 13, 236–251. doi:10.1109/TITB.2008.2009441.

Tosteson, A.N., Fryback, D.G., Hammond, C.S., Hanna, L.G., Grove, M.R., Brown, M., Wang, Q., Lindfors, K., Pisano, E.D., 2014. Consequences of false-positive screening mammograms. JAMA internal medicine 174, 954–961. doi:10.1001/jamainternmed.2014.981.

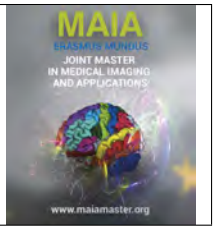
Wang, J., Yang, X., Cai, H., Tan, W., Jin, C., Li, L., 2016a. Discrimination of breast cancer with microcalcifications on mammography by deep learning. Scientific reports 6, 27327.

Wang, X., et al., 2016b. Deep learning in object recognition, detection, and segmentation. Foundations and Trends® in Signal Processing 8, 217–382. doi:http://dx.doi.org/10.1561/2000000071.

Zhu, W., Lou, Q., Vang, Y.S., Xie, X., 2017. Deep multi-instance networks with sparse label assignment for whole mammogram classification, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 603–611.

Medical Imaging and Applications

Master Thesis, June 2018



Brain Extraction on MP2Rage Images

Yuliia Kamkova, Pierrick Bourgeat

The Australian eHealth Research Centre, CSIRO Health and Biosecurity, Herston, Queensland, Australia

Abstract

Magnetization-Prepared 2 Rapid Gradient Echo (MP2Rage) is a new T1 weighted sequence in Magnetic Resonance Imaging (MRI). MP2Rage can obtain homogeneous intensity throughout the brain structures with a high dynamic intensity range between them. However, due to the sequence novelty, no effective skull stripping methodology has been proposed for it. In this study we developed a deep learning method, trained on a *silver standard* dataset, to provide an accurate skull stripping to MP2Rage images. The *silver* standard mask was created using a combination of MP2Rage and Fluid-Attenuated Inversion (FLAIR) images. The qualitative results shows that our deep learning approach is able to produce an accurate skull stripping on the MP2Rage images. Quantitative comparison of segmentation results of the brain structures with FreeSurfer software shows that it reduces gray matter over segmentation by 2.19% in terms of tissue volume. The method was also applied to the publicly available dataset of MPRage images (LPBA40) which comes with manually segmented brain masks. A Dice score of 97.88 was obtained, which is the highest result compared with state-of-the-art methods for this dataset. The results suggest this method is effective for processing skull stripping of T1 weighted MR images for both MP2Rage and MPRage.

Keywords: MRI, MP2Rage, skull stripping, deep learning, multi modality, bias field, silver mask

1. Introduction

Magnetic Resonance Imaging (MRI) is a widely used imaging modality to observe and study brain structures. Analysis of brain MRI images is commonly used to diagnose brain diseases such as Alzheimer's disease, aneurysms, sclerosis, brain tumour, Huntington's disease, brain abscess and many others.

There is a variety of software designed to process MRI images. The most commonly used automatic tool in neuroscience for brain segmentation from MR images is FreeSurfer (Fischl (2012)). It is designed to work with T1 weighted MR images. Magnetization-Prepared Rapid Gradient Echo (MPRage) (Mugler and Brookeman (1990)) which is the most frequently used T1 weighted sequence for structural brain imaging. Despite its wide use, MPRage suffers from inhomogeneity due to the heterogeneity of the bias field (B1) (that causes intensity variations of the same tissue across the image region), and low contrast between certain brain structures. Magnetization-Prepares 2 Rapid Gradient Echo (MP2Rage) (Marques et al. (2010)) is a new sequence which was developed to acquire bias-free T1 weighted images with an improved contrast-to-noise ratio between white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF). MP2Rage is able to enhance the contrast between them due to the double acquisition approach. In addition, the combination of both images results in a bias free image. However, while the GM to WM contrast is typically improved compared to MPRage, MP2Rage suffers from a similarity in intensities between dura matter, medium size vessels, and gray matter tissues. This makes the process of segmenting gray matter from skull tissues difficult (Marques et al. (2010)).

In the majority of methods used for processing MR images, including FreeSurfer, the initial step is skull stripping (SS). It includes preliminary processing which isolate the brain from extra-cranial or non-cerebral tissues (Kalavathi and Prasath (2016)). This differentiates white matter, gray matter and cerebrospinal fluid from the dura matter, sinuses, and all upper layers of the skull (Figure 1 from the Lubopitko Encyclopedia (http://encyclopedia.lubopitko-bg.com)). Most of the

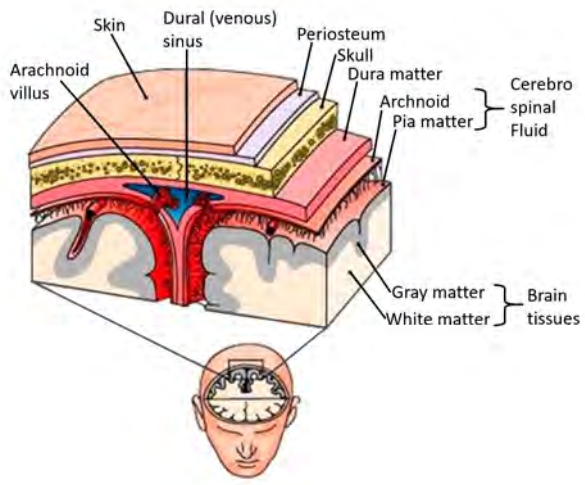


Figure 1: Brain skull morphology

tools are quite sensitive to the quality of the extracted brain mask. Therefore errors in this step can negatively affect downstream processing steps.

Considering the limitation of current segmentation algorithms for dealing with MP2Rage sequence, the main goal of this research was to create an efficient and robust method to perform skull stripping on MP2Rage images. Currently, state-of-the-art methods for image segmentation rely on use of deep learning (DL) (Litjens et al. (2017)). However, despite the good results achieved with this technique, DL requires a considerable amount of annotated data for training. Due to the novelty of the MP2Rage sequence, no manually annotated data, referred to as *gold standard*, is available.

Lucena et al. (2018) proposed a way to deal with the lack of the manual annotated data by creating a *silverstandard* segmentation mask. The set of segmentation masks were created using different automatic SS tools. The *silver* mask was then created by combining masks using simultaneous truth and performance level estimation (STAPLE) algorithm (Warfield et al. (2004)). While this method was successful in MPRage, it cannot be transferred to MP2Rage as the existing SS methods produce inaccurate results. For this reason, a different approach to define the *silverstandard* map was developed. Recently Viviani et al. (2017) presented that the combination of T1 weighted images with Fluid-Attenuated Inversion (FLAIR) can significantly aid in separating the dura from the brain tissues. Our methodology is to use the SPM (Statistical Parametric Mapping) segmentation toolkit to segment the brain tissues using both FLAIR and MP2Rage images. The final *silver standard* mask was created by combining information from SPM and a separate Expectation Maximization (EM) (Dellaert (2002)) algorithms. Using these masks, the DL approach was trained so that it can reliably segment MP2Rage even when no FLAIR data is available. Additionally, was compared to state of the art methods on an openly available dataset of MPRage images. Furthermore, the results was compared with the widely known FreeSurfer method.

2. State of the art

Existing methods for skull stripping can be divided into three main groups:

1. Manual segmentation performed by an expert, usually by a radiologist or other specialists in the field.
2. Segmentation performed by automatic skull stripping methods without any prior training procedure.
3. Segmentation methods based on deep learning, requiring training data.

The first method is extremely time-consuming, requiring more than an hour per volume. This makes this method expensive and unfeasible for large-scale studies.

Leading methods in the second group are based on a combination of intensity, edge detection, atlases, level set/graph cuts and registration tools (Galdamesa et al. (2011)). Due to the combination of different approaches, these methods can be time-consuming. Some of these methods also rely on the registration between images which imposes strong assumptions about the geometry, orientation, and features. Typically, most of the methods require manually tuned parameters which do not generalise well when using difference acquisitions parameters or imaging sequence (such as MP2Rage). The most widely used software for automatic brain segmentation is FreeSurfer by Fischl (2012). It is a publicly available open source software for the segmentation, meshing, cortical thickness estimation and statistical analysis of brain MR images. For SS it uses the Hybrid watershed algorithm (HWA) (Ségonne et al. (2004)), which is based on combining the watershed algorithm (Hahn and Peitgen (2000)) and deformable surface model (Dale et al. (1999)). Although this method provides accurate segmentation, the pipeline to perform segmentation is extremely time-consuming (approximately 17 hours per volume). Even though, it is a reference for the majority of novel methods in brain segmentation, which are striving to improve in terms of time per volume while showing similar accuracy.

SPM12 is the second most commonly used software in MRI brain tissue segmentation and analysis after FreeSurfer. It is a software package for the MatLAB environment that was developed by the researchers of Functional Imaging Laboratory of Welcome Department of Imaging Neuroscience at University College London (Ashburner et al. (2012)). For tissue segmentation, it uses a unified segmentation scheme (Ashburner and Friston (2005)), that relies on the spatial correspondence of pixels to the probability atlas and intensity distributions within 6 classes. As a result, it performs probability maps of the 6 structural groups of the brain (WM,

GM, CSF, 2 groups of skull tissues and background). This tool has been proven to be the epitome in segmenting CSF from WM and GM ((Tudorascu et al. (2016)). Moreover, this software can take information from multiple image modalities to create the tissue probability map.

Others automatic methods in skull stripping without any prior training are Brain Extraction Tool (BET) (Smith (2002)) from FSL toolkit (Jenkinson et al. (2012)), Advances Normalization Tool (ANTs) (Avants et al. (2009)), and Robust Learning Extraction(ROBEX) (Iglesias et al. (2011)). BET relies on the intensity-based threshold estimation between brain and non-brain tissues. It uses a deformable model that expands the sphere the centre of gravity until it reaches the edges of the brain. ANTs is based on the image registration with the template, deforming the brain mask to the subject space, which fuses them together using joint label fusion. ROBEX method also involves the registration of the subject to the template however the brain mask is generated using a random forest classifier.

The third group is the newest in the SS field. The advantage of using DL is that it can self-generate the features used to segment each of the tissues. However the disadvantages of this method are lack of generalisability of the approach, the computation requirement for training the network, and the need for a large training dataset. For each image from a new sequence or with different acquisition parameters, the network requires additional training with appropriate dataset. Within the DL frameworks for image segmentation there are two types of networks. One for skull stripping and one for brain segmentation using provided brain masks. Currently the leading skull stripping methods are CONSNet (Lucena et al. (2018)) and Deep 3D CNN (Kleesiek et al. (2016)). Both networks are using 2D Fully Convolutional Neural Networks (FCNN) U-net (Ronneberger et al. (2015)) based network architecture. Remarkably, CONSNet is able to generalise through the use of *silver standard* masks during the training. For brain tissues and structures segmentation outstanding results are obtained by 3D FCNN network (Dolz et al. (2017)), DeepNat - 2D FCNN U-net (Wachinger et al. (2017)), and patch based 3D FCNN (Bernal et al. (2018)). The V-net network (Milletari et al. (2016)) is a modified 3D U-net with added residual connections within each step of the encoding / decoding branches. The use of residual connections in the architecture provides a significant improvement in the performance of deep network, such as ResNet (He et al. (2016)). Adding residual connections to the network for brain segmentation could provide stability of the network and faster convergence. Bernal et al. (2018) has shown that using 3D patches with 3D convolutional filters gives the best result for brain segmentation .

3. Materials and methods

3.1. Materials

This investigation was conducted on the dataset of MP2Rage images. This sequence is a new approach in T1 weighted MRI data acquisition. One of the issue with MPRage is that it not only measures T1 relaxation but also M_0 (often referred to as proton density) and T_2, which contribute into the variability in intensity and contrast. MP2Rage images solve this issue by taking two images at different inversion times (GRE_{TI1} and GRE_{TI2}), but with identical sequence parameters. This results in the equal impact of B_1^-, M_0 and T_2^*. Combining the images by means of the ratio (Eq.1) will eliminate the impact of these parameters (Van de Moortele et al. (2009)) and create a bias free image (Figure 2-c)).

$$MP2Rage = \frac{GRE_{TI1} * GRE_{TI2}}{GRE_{TI1}^2 + GRE_{TI1}^2} \tag{1}$$

However, dividing the intensities with very small value in the background results in "salt and pepper" background noise. OBrien et al. (2013) proposed a simple way to denoise MP2Rage images (Eq. 2), by introducing the variable γ into each ratio.

$$MP2Rage = \frac{GRE_{TI1} * GRE_{TI2} - \gamma}{GRE_{TI1}^2 + GRE_{TI1}^2 + 2\gamma} \tag{2}$$

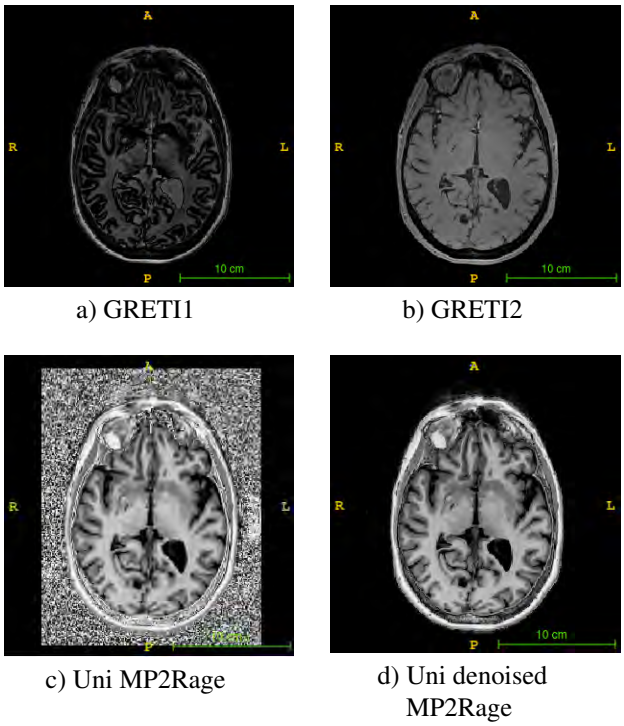
Introducing γ suppresses the noise presented on the uni MP2Rage image (Figure 2-d). However this is coupled with a small bias field that is slowly varying across the brain and requires further corrections.

The provided dataset also had FLAIR images, which is a T2 weighted MR image typically used to image white matter lesions. FLAIR provides clear contrast between dura and cortical gray matter. Including this sequence can assist when creating a robust brain mask and mitigate the over-segmentation presented in regions where dura matter and GM have similar pixel intensities on MP2Rage.

3.1.1. Imaging protocols

The data for the experiments was acquired as part of the prospective imaging study of aging (PISA). Images were acquired using a SIEMENS MAGNETOM Prisma machine. The study included 58 subjects. MP2Rage images were acquired with isotropic voxel size 1 mm, FOV 256 × 240, 192 sagittal slice, phase encoding anterior-posterior, phase oversampling = 10 %, TR = 5000 ms, TE = 2,96 ms, TI1 = 701 ms, TI2 = 2500 ms, non-selective inversion recovery, flip angle 1 =4°, flip angle 2 = 5°. Filters applied: distortion Correction (3D), Prescan normalize, acquisition time = 9 min 2s.

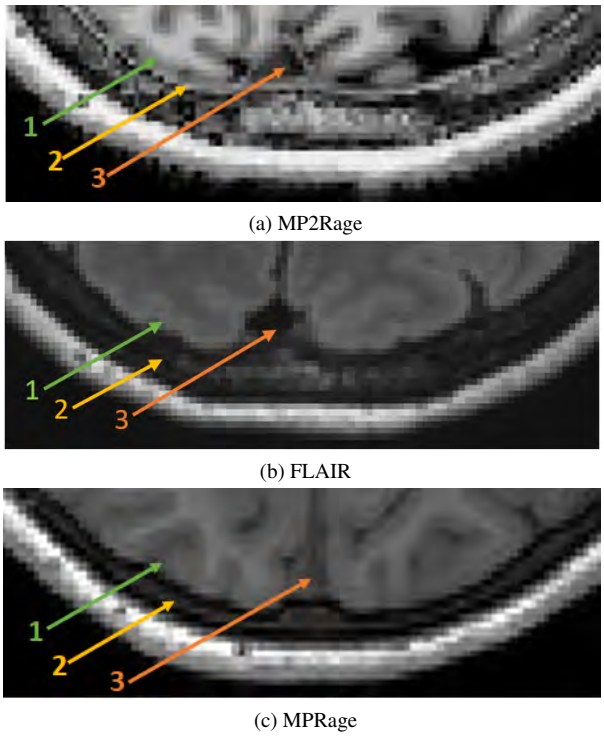
FLAIR images were acquired with the same dimension parameters as MP2Rage and phase encoding, without phase oversampling, TR = 5000 ms, TE = 388 ms, TI1 = 1800 ms, non-selective T2-IR inversion recovery.



a) GRETI1 b) GRETI2

c) Uni MP2Rage d) Uni denoised MP2Rage

Figure 2: MP2Rage sequence to obtain denoised T1 weighted MR images



(a) MP2Rage

(b) FLAIR

(c) MPRage

Figure 3: The difference in the representing internal structures of the brain within MP2Rage, FLAIR and MPRage (1 - gray matter, 2 - dura matter, 3 - superior sagittal sinus)

Filters applied: Raw filter, Distortion Correction(3D), Prescan normalize, acquisition time = 7 min 7s.

A second dataset LPBA40 which is a part of the LONI Probabilistic Brain Atlas project (http://www.loni.usc.edu/) was used to compare the developed approach with state-of-the-art methods in skull stripping. It includes 40 T1 weighted MPRage scans of healthy subjects with provided manual segmentation of the three tissues types inside the brain.

14 MP2Rages acquired from SIEMENS-MR scanner from a separate study was used as a third dataset. MP2Rage images are isotropic with voxel size 1 mm, FOV 240 × 256, TR = 5000 ms, TE = 2,96 ms, TI1 = 0ms, TI2 = 0 ms, acquisition time 8 min 18s. No FLAIR images or segmentation were provided for this study.

3.2. Methods

The networks applied in this study were implemented using Python 3, using TensorFlow libraries (https://www.tensorflow.org/). Both networks were trained using the *silver* mask to predict segmentation for MP2Rage image.

Figure 4 shows steps to create the *silver standard* segmentation mask. It was implemented using SPM software package (version 12, Welcome Trust Centre for Neuroimaging, University College London, http://www.fil.ion.ucl.ac.uk/spm/) from MatLab (version 2016b, The MathWorks, Inc.) and FSL toolbox (FMRIB Software Library, Release 5.0 (c) 2012, The University of Oxford).

FreeSurfer toolkit (version 5, Laboratory for Computational Neuroimaging at the Athinoula A. Martinos Center for Biomedical Imaging, https://surfer.nmr.mgh.harvard.edu/) was used to compare quantitative and qualitative differences between segmentation.

3.3. Creating Silver standard mask

3.3.1. Register FLAIR images to MP2Rage uni image

SPM requires primary registration of all input modalities. In the PISA study, MP2Rage images were obtained with corresponding FLAIR images for each of the subject. Despite this, registration of FLAIR to MP2Rage is still required due to the movement of the subject between image acquisition. Due to absence of shape variability for one subject, only linear registration was applied (translation and rotation). We used ”FLIRT linear registration” (Jenkinson and Smith (2001), Jenkinson et al. (2002)) from the FSL toolbox.

Qualatative inspection of the segmentations was performed using ITK-Snap (version 3.6.0). It is a powerful tool for 3D image visualisation that allows to overlay the segmentation on the image.

3.3.2. Tight brain mask using SPM segmentation

The brain mask for each of the volume was created using the Unified Segmentation procedure from the

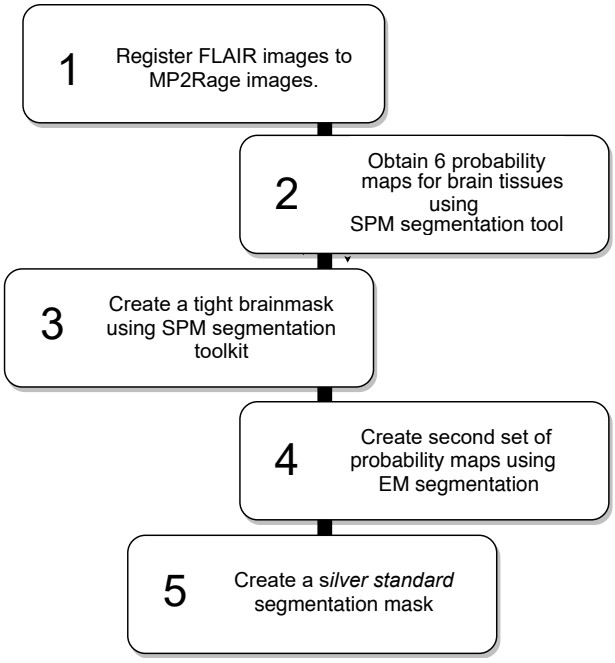


Figure 4: Workflow of the presented paper

SPM software package (Ashburner and Friston (2005)). The MP2Rage was supplied as a first input channel and the FLAIR image as a second channel for the multispectral classification. The parameters of the program were also adjusted according to minimum representation of B1 on MP2rage and FLAIR. The number of Gaussians used to represent the intensity distribution of each tissue type was tuned to include different intensity variations. For white matter and gray matter two Gaussians were used, while only one was used for CSF.

Skull stripping was conducted using the generated probability maps of WM, GM, and CSF. The probability maps were then converted to binary masks, with thresholds of 0.1, 0.75 and 0.9 for the WM, GM, and CSF respectively. The threshold for each of the tissue type was chosen empirically to ensure that dura matter is not included in the brain mask. Afterwards, tight skull stripping mask was created using morphological operations on the merged masks from three tissues segmentations.

3.3.3. Refining SPM segmentation

To perform brain tissue segmentation, a combination of probability maps was used. The first map was derived from previous steps and the second map was generated using the EM algorithm. The EM algorithm is a well-known statistical tool for creating probability maps based on the intensity distribution of pixels through classes. It takes into account both image modalities (FLAIR and MP2Rage) and their correspondences between pixels. Unified segmentation combines the EM algorithm for 6 tissues from the whole scan with the spacing information of the pixels. Whereas the second probability map is based only on the intensity distribution inside the three classes (WM, GM, and CSF) of the brain. The final segmentation was conducted by following a maximum-a-posteriori approach over the combined probability maps $P1(X) * P2(X)$, where $P1(X)$ and $P2(X)$ are the probability maps obtained with the Unified Segmentation and EM methods respectively. For pixels inside the brain that have zero value in the SPM probability maps, the segmentation based only on the intensity information (EM probability maps) was applied.

3.4. Segmentation using Deep Learning

To perform the brain segmentation using deep learning, we used two deep network architectures, both of them representing 3D modification of the standard U-net.

Within the DL frameworks for image segmentation, there are two main approaches: a) voxel wise, which extracts the patches for each of the pixel and predicts the label individually b) fully convolutional, which interprets the full image and predicts the label for entire subject in one feed-forward step. While the second approach is much faster compared to the first one, it requires a lot more memory to store the features and training data. Patches can be extracted either in 2D, 2.5D (when extracting three planar anatomical 2D patches around one pixel) or 3D. A new tri-planar approach was introduced by Lucena et al. (2018) for skull stripping. This approach utilises randomly 2D patches extracted from each of the anatomical view. The difference between this method with a traditional 2.5D approach that 2D patches surround not one voxel but lay independently in all 3 dimensions. This method aims to imitate the *gold standard* segmentation. It gives promising results in skull stripping, however, there is no application in tissue segmentation, which is led us to apply 3D patches and 3D convolutions in our work.

The first architecture represents a 3D - Unet based network. It contains 4 blocks of encodes levels and 3 corresponding decoding blocks. Each block consists of 3 fully connected convolutional layers. In the encoding part, the number of convolutional filters (channels) is increased with each block, from 1 channel in the top layer to 128 channels in the lowest layer.

The original U-net was designed to work on full size images where the size of the input image guided the number of layers. Since our patches are smaller, this impacts the number of downsampling that can be perform. Due to this limitation, the number of layers was decreased from five to four. The amount of channels is only increased once within each block, whereas in the original architecture, the number of channels was increased multiple times per block. The concatenation of the feature map from the encoding path ensures the information from high-resolution features has been taken into account.

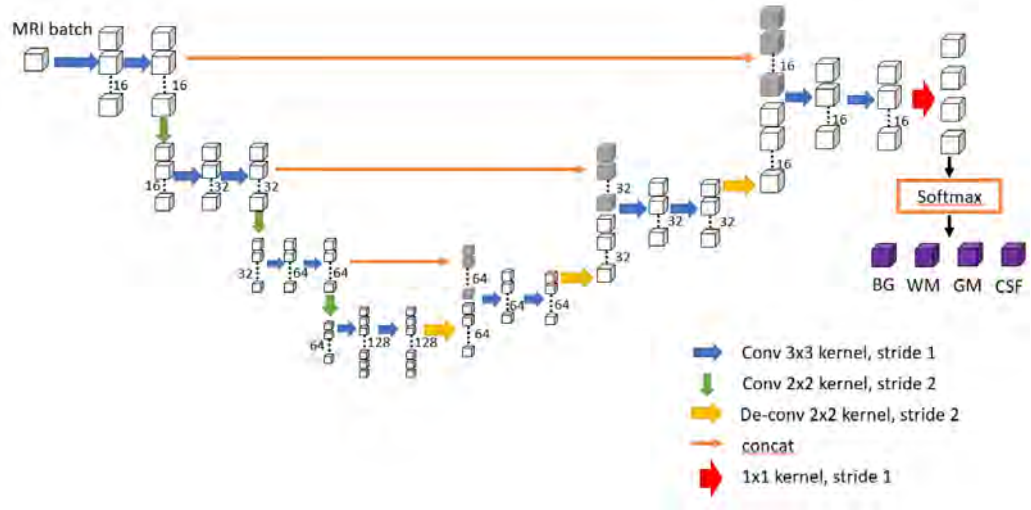


Figure 5: U-net based architecture

The second network represents a V-net architecture, another 3D modification of the U-net. The main difference from the 3D U-net is the residual connections which perform an element-wise summation within a block of convolution filters of the input feature map with the output feature map. As for the U-net, the number of layers is different from the original V-net architecture. Unlike the previous approach, the increasing number of channels on the compression path occurs in the downsampling stage, by implementing filters of 2x2x2 and appropriate padding. The batch of 50 3D isotropic patches was used as an input for the network. Bernal et al. (2018) shown that overlapping of patches can increase the performance of segmentation. Therefore, an overlap of 50% in each dimension was used to train the network with all possible positions of the patch and increase the number of patches to train the network. In the reconstruction step the mean between all probabilities between patches was calculated. The patches were first normalised to zero mean, unit variance coupled with a selective filter to exclude patches that do not contain any information from the extracted patches. Intensity and magnitude gradient thresholds were applied to exclude patches that belonged to the background and do not contain strong edges.

Both architectures represent FCNN. In comparison with different traditional networks, such as GoogleNet or AlexNet, FCNN does not have limitations on the input size of the image because the whole network could be presented as single non-linear convolution, trained from end-to-end. FCNN includes spatial information by obtaining the feature map in one dense inference step making the computations more efficient. This mitigate against redundant convolutions and pooling in the network architecture.

For both networks, each convolution layer was followed by PreLu activation function. It was shown that using PreLu as activation function improves the performance of the network (Dolz et al. (2017)). Adam optimiser (Kingma and Ba (2014)) was used in both networks. One of the characteristic of this optimiser is that it performs better without maxpooling (Kingma and Ba (2014)). Therefore, maxpooling was omitted in the presented networks. In both networks batch renormalisation proposed (Ioffe (2017)) was used. Traditional batch normalization (Ioffe and Szegedy (2015))increases the stability of a neural network by normalizing the output of each previous activation layer. Batch renomlization is an extension of the traditional batch normalization. It ensures that batches are normalized on individual examples rather than the entire batch. The loss function for both networks was the general Dice coefficient loss. It was calculated as a mean value of the Dice coefficient for all classes that we aimed to segment. The goal of the networks was to maximize this value. For both networks, a learning rate of 0.01 was used. To reduce the number of channels at the final layer from 16 to the number of predicted classes (4 channels in the presented problem), a convolution filter 1 x 1 x 1 was used. A softmax function was applied to create four probability maps corresponding to the 4 classes(Background(BG), WM, GM, and CSF) for each patch of the network. For the final reconstruction, the reconstruction was computed for each pixel individually. The mean of all probability patches that were created for the pixels were compared. The class with the highest probability was predicted as a final label.

3.5. Validation

Three metrics were used to evaluate the quality of the segmentation:

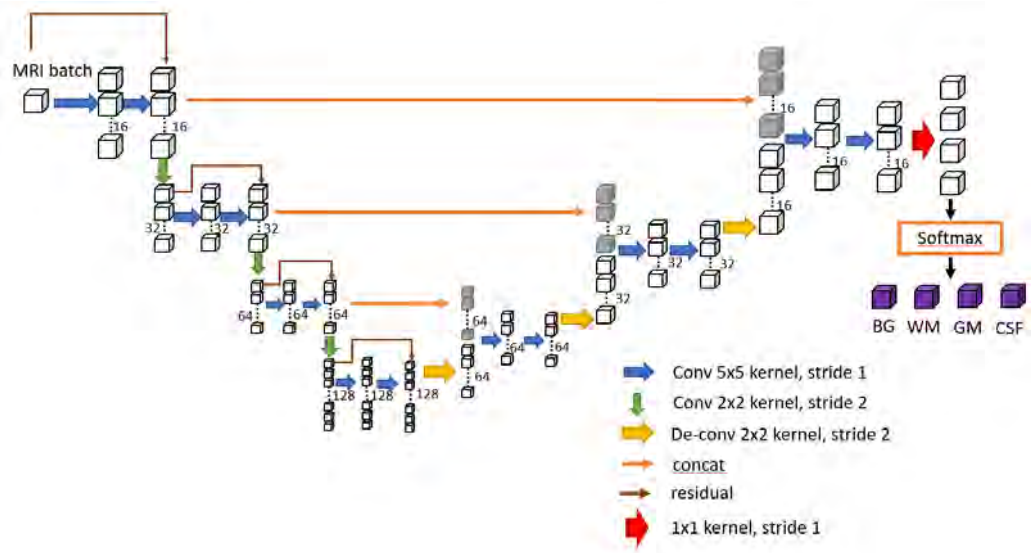


Figure 6: V-net based architecture

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{3}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{5}$$

Where TP, FP, TN and FN are a number of true positives, false positives, true negatives and false negatives respectively. Sensitivity measures the percentage of correctly segmented brain tissues inside the brain while specificity presents the percentage of correctly segmented non-brain tissues. The Dice coefficient metric is a compromise between sensitivity and specificity, it evaluates trade-off between the correct and false voxel predictions. All of the presented metrics are designed to work with only two classes.

In cases where 4 classes are presented, Dice coefficients were calculated separately for each of them (excluding the background) and their mean was calculated. For tuning the network parameters 3 different Dice coefficients were evaluated:

- mean value of the Dice score to the 3 tissues segmentation
- Dice score of the gray matter only
- Dice score of the whole skull segmentation

PISA dataset was divided into three subsets for the experiments: 60% training, 20% validation, and 20% testing set. Tuning of the network was performed on the validation set. The test set was applied in the final step selecting the best network within U-net and V-net for skull stripping in the provided dataset.

3.6. Generalisability

The second dataset containing MP2Rage images to ensure the robustness of the network. The images were acquired with different scanners and different parameters compared to the PISA study. To reduce the bias between the 2 datasets, two pre-processing steps were evaluated: 1) matching the histograms of the new dataset to the one that the networks was trained on 2) normalizing the PISA dataset in the range from 0-255, training train the network on this data, matching the new dataset to the training data to obtain the final segmentation.

3.7. Comparison with State-of-the-art

To evaluate the performance of the network for skull stripping on MP2Rage, the dataset LPBA40 was used. It is a common dataset to compare the performance of skull stripping methods, which contains T1 weighted MPRage images. Since MPRage is quite different from MP2Rage in terms of contrast and pixel intensities between structures, the network was trained from scratch using LPBA40 data. However no tuning for this dataset was done. The parameters were taken from the network for PISA segmentation networks. To have an accurate result, the validation was performed using a three-fold cross validations. The network was trained to segment 4 classes in one step (WM, GW, CSF, and BG). In post-processing, the brain mask was created by merging the 3 classes that belong to the brain and compared with other methods.

3.8. Segmentation using FreeSurfer

In Freesurfer, the segmentation task is performed in three stages (Figure 7). At the end of the first step of FreeSurfer pipeline, 5 volumes are available:

orig.mgz (motion correction, image normalization between 0−255 and size correction), nu.mgz (non-uniform normalization), T1.mgz (intensity normalization with wm equal to 110), and brainmask.mgz (that represents extracted brain from the skull).

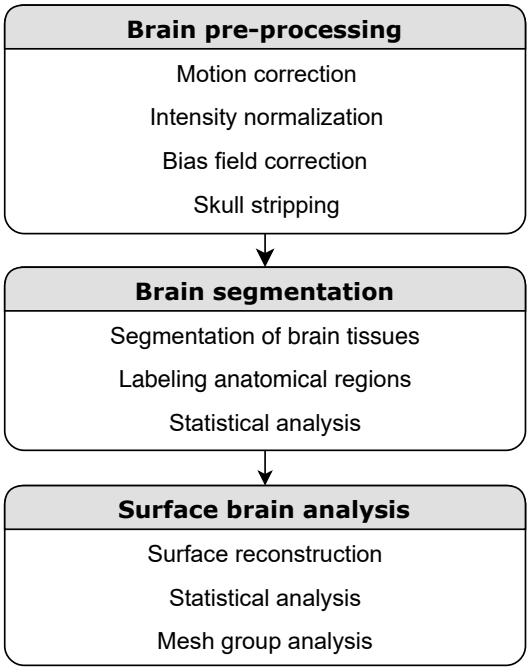


Figure 7: Workflow of FreeSurfer

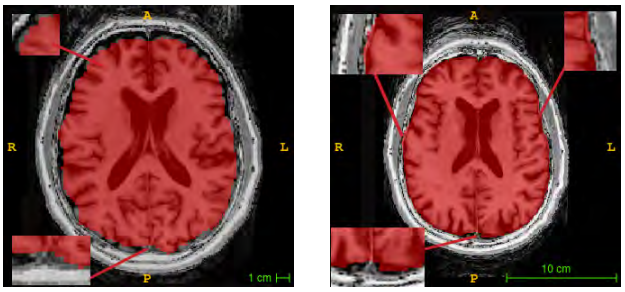
At the end of the first stage, the brain mask can be replaced with the one that we generated using our deep learning method. The rest of the pipeline of FreeSurfer can then be run using this brain mask. The whole pipeline was also run without any modification, to compare the results with and without our automatic SS.

The most commonly obtained error observed on the FreeSurfer brain masks was the over-segmentation of the brain in the borders of the dura and gray matter, due to the similarity of intensities. To quantitatively measure the improvement with the new mask, we compared the measures of the thickness of the brain matters and structures inside the brain. If the mask is accurate and does not contain over segmentation, the thickness of the gray matter is expected to be smaller using the new mask.

4. Results

To create the brain mask for the MP2Rage images were explored 4 automatic skull stripping tools that do not use any prior training. Figure 8 presented results of applying 2 methods that are commonly used in the MPRage sequence for skull stripping.

The masks produced with ROBEX method cuts the GM and CSF on the borders with the brain resulting in overall under segmentation. The mask from ANT's segmentation shows over segmenting in the borders where



a) ROBEX b) ANT's

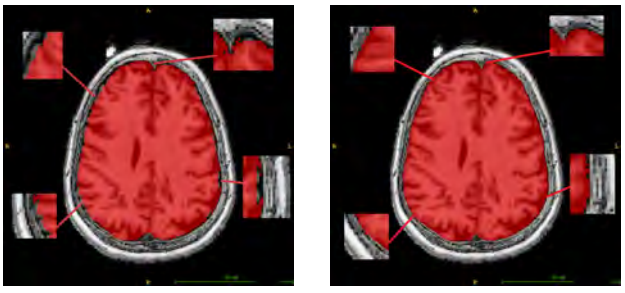
Figure 8: Skull stripping on the MP2Rage using automatic brain segmentation tools without training; in red - masks obtained with different methods

GM is touching the dura matter. In both cases the CSF around venous sinus is not segmented correctly.

Aside from the presented method, the results from BET segmentation tool was tested. The segmentation produced by BET include the skull bone as a brain tissue. As the goal of this study was not the comparison of the SS tools on the MP2Rage, the images of this tool performance is not included.

4.1. Creating silver standard segmentation

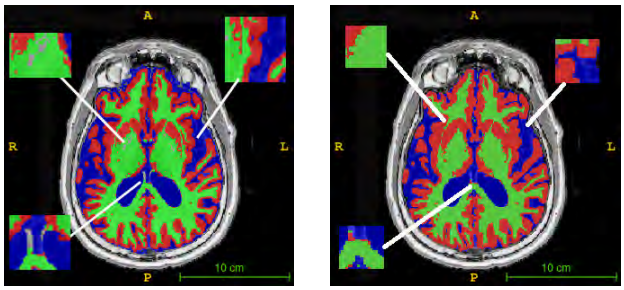
Using FLAIR images significantly improved the qualitative performance of the SPM segmentation. However after running the data in the toolbox with standard parameters (two Gaussians for the CSF and one for WM and GM), an over-segmentation of the gray matter was observed. The number of Gaussian distributions impacts on the quality of SPM segmentation for each of the tissue type. After multiple experiments, it was decided to use one Gaussian distribution for CSF and two for GM and WM. The differences between the results of the segmentation with higher number of Gaussians for WM and GM was not significant, but the time for processing each volume increased dramatically (twice per one additional distribution). For this reason, it was decided not use more than 2 Gaussians.



a) MP2Rage b) MP2Rage + FLAIR

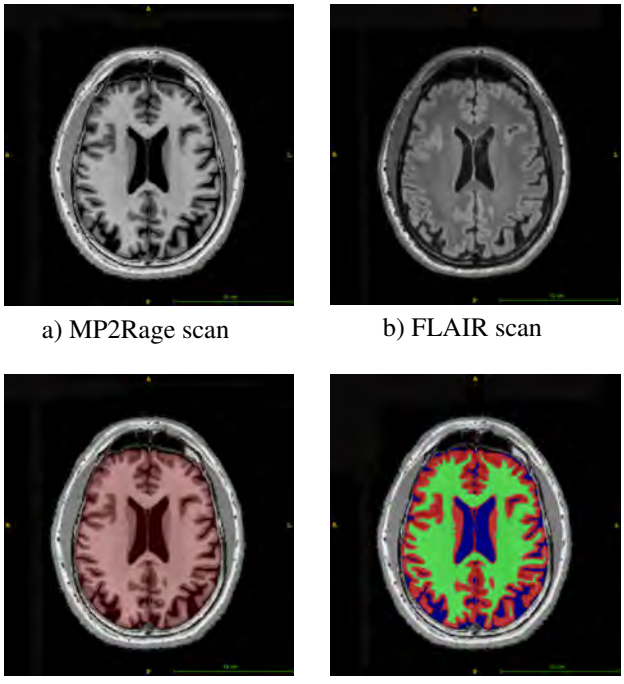
Figure 9: Skull stripping on the MP2Rage using SPM and different image modalities

The time per image in SPM toolkit with the parameters mentioned above was about 15 min 40s. An example brain mask using FLAIR and MP2Rage is presented in the Figure **??**. For creating the WM, GM and CSF segmentation masks the SPM probability maps inside the brain mask were combined with the probability mask from the EM algorithm. Figure 10 illustrates the results of combining the two probability maps, creating an accurate segmentation.



a) Unified segmentation b) Unified + EM segmentation

Figure 10: Brain tissue segmentation using different probability maps



a) MP2Rage scan b) FLAIR scan

c) Skull stripping mask d) Segmentation of GM, WM, CSF

Figure 11: Creating the ground truth for tissue segmentation

Figure 11 illustrates how a multi-modal approach can create an accurate brain mask. The venous sinuses are correctly segmented from the dura matter. Visual inspection did not reveal any significant over-segmenting of the grey matter in the brain mask. Overall, qualitative inspection gave satisfying results.

4.2. Segmentation using Deep Learning

The tuning of the network parameters was performed on the validation set of the PISA MP2Rage dataset. The parameters that were optimized are the patch size, batch size, learning rate, batch normalisation, thresholds for the patch intensity and magnitude of the gradient.

To define the patch size to use for the network, it was trained using different isotropic patches size. The results evaluated in terms of Dice score for the three tissues (Table 1), grey matter (Table 2) an the whole brain segmentation (Table 3) were checked. The tested size were limited to 12 pixels on the lower end as smaller size would risk excluding contextual information, and 50 pixels on the higher end because of memory limitations.

Table 1: Mean Dice coefficient through GM, WM and CSF for the whole scan in the validation dataset PISA study

Patch size	12	**32**	50
Dice (V-net)	90.95±2.46	**92.46 ± 1.47**	92.22± 1.33
Dice (U-net)	91.87 ± 1.48	**91.89 ± 2.16**	91.73 ±1.73

Table 1 shows that the best Dice value was obtained for the patch of 32×32×32 for both networks. V-net shows better Dice value and standard deviation for patches of 32×32×32 and 50×50×50, when for the 12×12×12 U-net performs higher.

Table 2: Mean Dice coefficient through GM for the whole scan in the validation dataset PISA study

Patch size	12	32	**50**
Dice (V-net)	92.02±2.18	93.53 ± 1.34	**93.54± 1.28**
Dice (U-net)	93.60 ± 2.01	92.47 ± 3.50	**92.96 ±1.41**

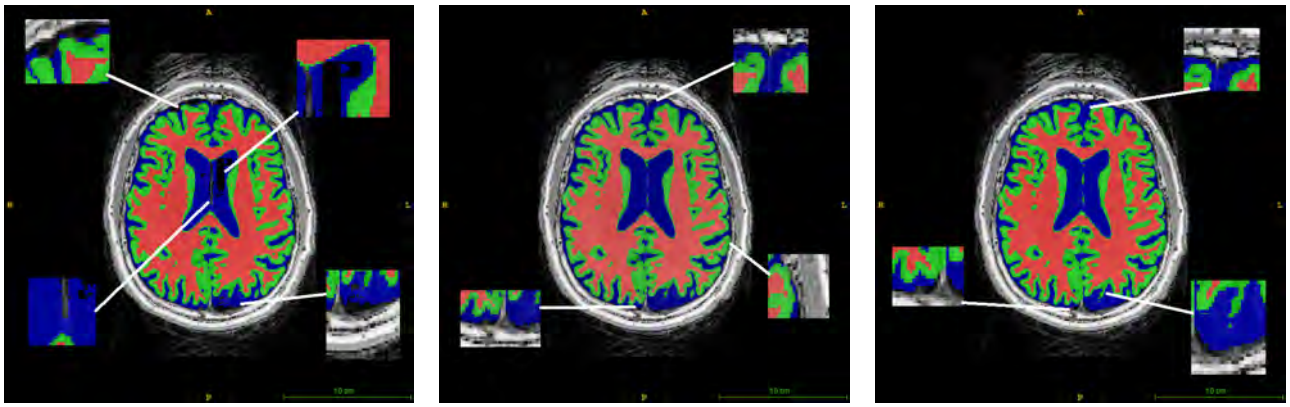
In terms of the Gray matter segmentation (Table 2), the network trained with patches of 50×50×50 has 0.01 higher Dice value compare to the of 32×32×32 patch. Comparing between the networks, the performance of the V-net is higher for these patches which is similar to the mean Dice score (Table 1).

Table 3: Mean Dice coefficient through brain segmentation for the whole scan in the validation dataset PISA study

Patch size	12	32	**50**
Dice (V-net)	98.32±0.44	98.63 ±0.30	**98.63±0.29**
Dice (U-net)	98.60 ±0.33	98.23 ±1.57	**98.54 ±0.58**

Table 3 presented the performance the networks trained with different patch sizes for the skull stripping. For both patches of 32×32×32 and 50×50×50 the performance of V-net is equal where as the U-net had better performance with the patch of 50×50×50.

Qualitative performance of the networks are presented in the Figure 12.



a) Patch 12x12x12 b) Patch 32x32x32 c) Patch 50x50x50

Figure 12: Mask generation using different patches

In the Figure 12 it is observed that the reconstruction with the patch 12 ×12×12 has errors in classifications between the tissues inside the brain with BG. When on the reconstruction with big patches the absence of small details was highlighted.

Tuning of the networks parameters was applied on patches of size 32 ×32×32 with the V-net network. The evaluation metric was the mean Dice coefficient computed across all patches of the validation set at the best epoch out of 40.

The threshold mean was the first parameter that defines cut-off on the mean patch intensity and is used to exclude background patches. A percentile of the mean patch intensity is compared to the whole scan intensity data. If mean value of all patch voxels is less than TM, then the patch will be excluded. This argument was implemented to include patches of brain scans that had weak edges, but moderate intensity (e.g. homogeneous patch of white and/or grey matter).

Table 4: Dependence of the Dice coefficient from the threshold mean value

TM, %	30	35	40	45
Dice	92.7 ± 8.55	**92.71 ± 8.65**	92.37 ± 7.98	92.34 ± 8.34

Table 4 shows that the best Dice index obtained with the value of TM of 35%. A higher value leads to smaller Dice, no difference is observed for smaller values.

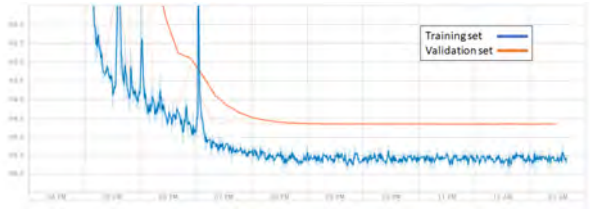
The threshold gradient-magnitude (TGM) defines the cut-off on the gradient magnitude image and is used to exclude patches with weak edges (e.g. homogeneous tissue outside the brain, blank background). Similarly to TM, the threshold was defined as the percentage of the mean intensity within the patch of the gradient image compared to the mean intensity within the full gradient image. The patch will be excluded if the gradient is lower than the value of TGM. This gradient-magnitude testing was implemented to reject patches with weak edges (e.g. homogeneous tissue outside the brain, blank background).

Table 5: Dependence of the Dice coefficient from the threshold gradient magnitude value

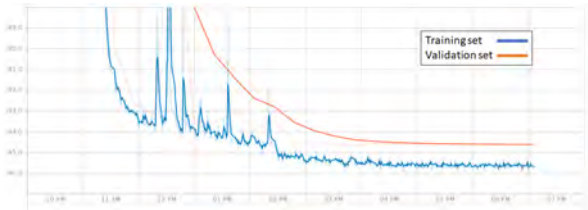
TGM, %	60	70	80	90
Dice	92.75 ± 8.73	**92.86 ± 8.63**	92.70 ± 8.85	92.67 ± 8.76

Table 5 shows that the best value of Dice achieved using TGM = 70%.

Figure 13 presented the variation of the Dice coefficient loss during the training time for segmenting brain tissues. Both networks were trained in 40 epochs. The best validation loss for the V-net was archived on the 16 epoch, for the U-net on the 12 epoch.



(a) U-net Dice loss in 40 epochs



(b) V-net Dice loss in 40 epochs

Figure 13: Training and validation loss for the U-net and V-net architectures, patch 32 × 32 × 32, 40 epochs

Figure 13 shows that for networks with such architectures, the learning process converges conducts after the first 10-15 epochs.

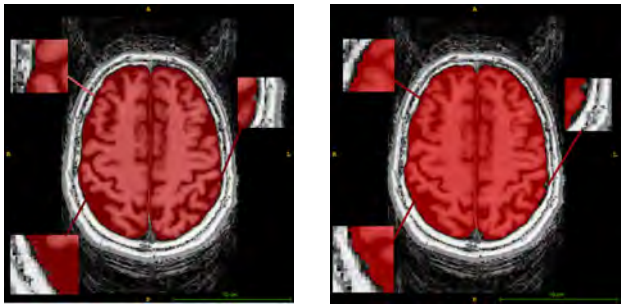
The Dice coefficient on the testing set is presented for both tuned networks in Table 6. The Dice coefficient was calculated for the whole reconstructed volume. The test set contains 11 volumes, that were completely new for the network, as they were not used in any of the previous tuning. Due to these reasons, no cross fold validation was performed. This set was used for checking the difference in FreeSurfer performance.

Table 6: The Dice coefficients for two networks in the reconstructed test set from the PISA study

Patch size	12	**32**	50
Dice(V-net)	90.95 ± 2.46	**92.46 ± 1.47**	92.22 ± 1.33
Dice(U-net)	91.86 ± 1.48	**91.89 ± 2.16**	91.73 ± 1.73

Table 6 shows that for the testing set, the highest Dice score was obtained using the V-net architecture. For all three tissues the performance of the V-net increased the Dice value by 0.12, whereas for the GM, the value was increased by 0.08. For the skull stripping V-net had the highest Dice coefficient, with the smallest standard deviation.

The qualitative comparison of the skull stripping using V-net and its corresponding *silver standard* is presented on Figure 14.



a) Deep learning mask b) Silver standard mask

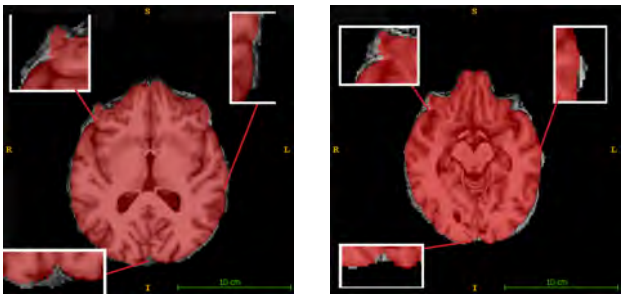
Figure 14: Comparison of the mask obtained with two different methods

As seen on Figure 14, the mask created with deep learning approach succeeded to improve the brain mask in areas where grey matter is in contact with dura matter. In addition some areas of CSF are segmented more accurately by the deep learning approach.

4.3. Segmentation using FreeSurfer

Figure 15 presents the overlap of the brainmasks generated with V-net deep learning method and with FreeSurfer for two different subjects.

In both subjects, FreeSurfer tends to over-segment the brain in places where the gray matter is in contact with dura matter. Using the new mask resolves this problem. In addition, venous sinuous are omitted from the brainmask.



a) b)

Figure 15: Extracted brain using two different masks for two different cases (in red colour - Deep Learning mask)

Figure 16 shows the cortical GM volume using the default FreeSurfer pipeline and the one using our DL brain mask. Using the DL mask resulted in thinner GM, likely due to the exclusion of the dura from the GM segmentation.

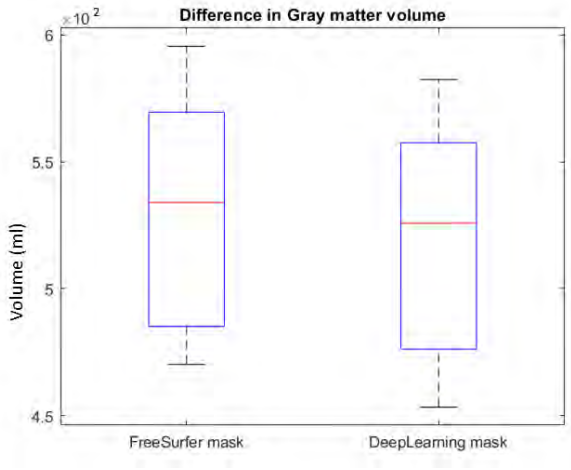


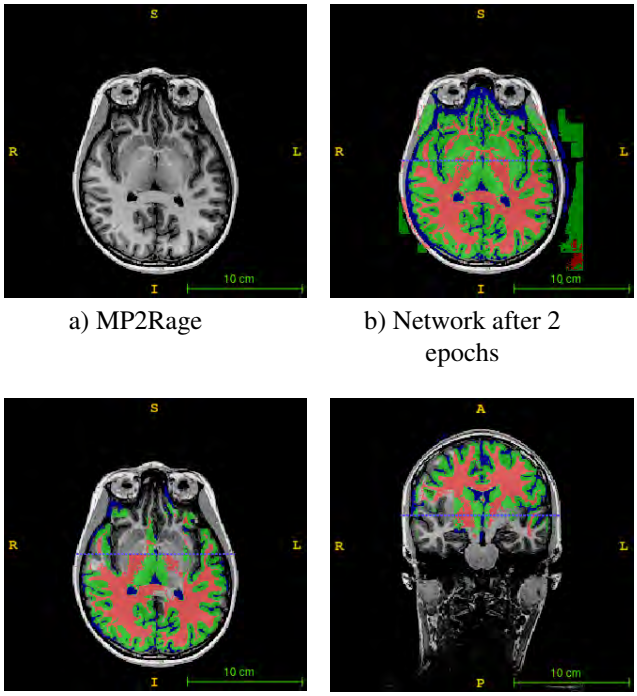
Figure 16: Gray matter thickness

Observed in Figure 16 that the volume of gray matter decreased from mean value of 529.9 ± 47.68*ml* to 518.29 ± 47.82*ml*, which made a difference in 2.19%. In addition, the difference in the CSF and WM volumetric thicknesses was measured, the improvement of 0.27% for CSF and 0.001% for WM using the mask from deep learning method was detected.

4.4. Generalisability of the network

The dataset of MP2Rage images from a different study was segmented using the network trained on the PISA dataset. Application of the network to the raw image data showed poor results. Pre-processing steps were carried out for matching the new dataset to the images, used for training the network. The first proposed method includes histogram matching of new dataset to the volume from the training set of PISA dataset. However, this approach did not show any qualitative im-

provement. The second experiment includes normalisation both datasets in the range [0−255]. Afterwards histogram matching was applied. Figure 17 illustrates the best segmentation results obtained after different epochs during the deep learning training process.



a) MP2Rage b) Network after 2 epochs

c) Fully trained network (40 epochs) d) Fully trained network (40 epochs)

Figure 17: Brain segmentation on the new study using the network trained on the images from PISA study

Figure 17 shows that using the network trained on the data from the PISA study poor results were obtained. Checking the segmentation from the different epochs of the networks shows that in the second epochs the network is able segment internal tissues, however the segmentation of the brain from the background failed. With the fully trained network was applied to the images, a majority of the internal tissues of the brain are classified as a background.

4.5. Comparison with state-of-the-art

The comparison of the skull stripping done by V-net with the state-of-the-art methods in the LPBA40 dataset is presented on the Table 7.

In Table 7, the patch based V-net has achieved the best performance for skull stripping in terms of Dice coefficient, improving the previous results from Lucena et al. (2018) for this dataset by 0.3 Dice. Also the U-net was tested for this dataset. The mean Dice through three cross validation was 97.87 ± 0.011 which also outperforms the previous results. The V-net ranks as the third in terms of the Sensitivity, that is higher than the two previously proposed deep learning approaches (3D CNN and CONSNet). In the Specificity comparison, the V-net is on the second rank, with a 0.12 difference compared to the leading 3D CNN and 0.16 higher that CONSNet.

Table 7: The performance of the state-of-the-art methods in skull stripping for the LPBA40 dataset

Method	Metrics		
	Dice (%)	Sensitivity (%)	Specificity ()%
BET	96.625 ± 0.007	97.236 ± 0.014	99.279 ± 0.002
HWA	92.515 ± 0.012	**99.898 ± 0.012**	97.092 ± 0.002
Deep 3D CNN	95.696 ± 0.007	92.614 ± 0.015	**99.831 ± 0.001**
CONSNet	97.353 ± 0.003	97.257 ± 0.007	99.541 ± 0.001
STAPLE	97.585 ± 0.002	98.144 ± 0.006	99.457 ± 0.002
Our V-net	**97.88 ± 0.008**	97.72 ± 0.008	99.71 ± 0.007

5. Discussion

The qualitative inspection of the brain masks produced by the state-of-the-art automatic tools (Figure 8) highlighted the errors in the segmentations. It should be noted that these methods are based on the image registration to the template (ANTs) or to the Atlas (ROBEX), which can be biased when the template is using a sequence different to that of the image. In both cases there were no available MP2Rage images.

Adding the FLAIR modality to the SPM segmentation tool improved the brainmask segmentation significantly. In those areas where vessels are close to the brain sinuses, or where gray matter is in contact with the dura matter, FLAIR modality has a better contrast in tissue intensities. As a result, the segmentation mask presented in Figure 11 does not includes any vessels and achieved to segment dura matter from the gray matter and CSF with greater clarity. The qualitative comparison of this mask with the outputs of other automatic methods found this approach to be the most suitable for performing MP2Rage skull stripping. Our results, in agreement with the ones obtained by (Viviani et al. (2017)), confirmed that combining the FLAIR images with T1 MR images could be used to performed accurate CSF and GM segmentation from dura matter.

Figure (10) presents the improvement in tissue segmentation achieved by adding the probability map created by EM to the obtained ones with SPM. The final achieved segmentations were used as a training *silver* mask in two deep learning convolutional networks. By means of this approach, it was tackled the the limitation of requiring multiple image modalities for creating the brain mask.

Three different patch size were evaluated in order to determine the input patch size that gave the most accurate results in our DL networks. Tables 1-3 presents the Dice score in the validation set. Table 1 showed that the best for performing tissue segmentation for both networks will be the with the 32x32x32 patch. However the next experiments shows that the patch of 50x50x50

performs the same Dice in terms of the gray matter segmentation and skull stripping (Table 2 and Table 3). Due to the not significant difference in Dice score (0.01 for GM and absence of it for SS for V-net, 0.5 for GM and 0.3 for SS for the U-net), we also took into account qualitative consideration of the reconstructed patches (Figure 12). The patches of 32x32x32 has a slightly difference on the segmentation of venous sinus segmentation from dura matter which is a common error in MP2Rage segmentation. Consequently, taking into account both quantitative and qualitative measures, the patches of 32x32x32 was chosen as an input to both networks.

The experiments with tuning the network in terms of patch selection for the input to the network showed that changing the parameters for extracting patches could improve the performance by 1 Dice score (Table 4-5).

We compared the performance of the V-net and U-net based architectures on the testing set, to evaluate the effect of the residual connections improve on the performance of the network (Table 6). For the patch of chosen size, the performance of V-net architecture on the test set improved the Dice value for GM, WM and CSF in 0.12 Dice compare to the U-net. In terms of skull stripping the V-net has also higher performance and better stability of Dice value. The residual connections are designed to stabilise and improve the convergence time of the network. However, while the Dice score overall is improved, the experiments showed it did not translate in an improved convergence speed, with with the U-net converging after 11 epochs, while the V-net achieved it after 14 epochs. Also, as seen in the Figure 13 V-net had more outliers compared to the U-net. Thus, *wasdoneanassumption*, that in the case of brain segmentation, the U-net architecture is quite stable itself and residual layers can create outliers, and slow down the rate of convergence. Qualitative inspection of the obtained segmentations from deep learning compared to the *silver standard* segmentation masks, concluded that the deep learning has improved the segmentation of the where gray matter is connected to dura mater (Figure 14). This is a common error conducted by automatic tools on the MP2Rage images. As a result deep learning method was implemented into the further experiments with FreeSurfer.

Comparison of brain mask obtained from FreeSurfer and using the deep learning method showed that deep learning method effectively mitigates over-segmentation of gray matter. (Figure 15). FreeSurfer statistical analysis was used to quantitatively compare the impact of the new mask to the GM, WM and CSF segmentations. Figure 16 shows that the biggest difference was achieved in the gray matter volume. The CSF volume also decreased, while the white matter stay almost unchanged. This confirmed the qualitative observation that the deep learning mask reduces the gray matter segmentations on the MP2Rage images.

When testing the network to separate dataset of MP2Rage images, poor results were obtained, with large misclassification of brain tissue into background. Despite pre-processing the images with histogram matching and scan normalisation, errors were observed in the areas with low intensity. This results shows the weak generalisability of the deep learning methods when difference acquisition parameters or scanners are used. Future work will focus on testing different pre-processing steps and on improving the future generelizability of the network overall.

The deep learning method was applied to the tissue segmentation of the publicly available dataset LPBA40 (Table 7). The brain mask was obtained by merging the manual tissue segmentations. The mask was compared with state-art-methods in terms of Dice coefficient, Specificity and Sensitivity using data publishes in Lucena et al. (2018). Two deep learning methods and 3 automatic methods were compared. The V-net approach showed the highest performance in terms of the Dice score, that shows the effectiveness of proposed approach in skull-stripping on MPRage images with available training data.

6. Conclusions

In this study, we developed an effective deep learning method to perform skull stripping to T1 weighted MR images. The network for MP2Rage images was trained using the *silver standard* mask generated by SPM unified segmentation using a combination of FLAIR and MP2Rage images. The obtained mask was able to successfully segment the dura mater from the gray matter and CSF tissues. The capability of the method for performing gray matter segmentation was quantitatively and qualitatively confirmed by comparing the obtained masks with the FreeSurfer obtained ones. However, the method did not show robustness for conducting the task over data with different acquisition parameters. On the MPRage dataset with available training data, the deep learning method obtained state-of-the-art results. This approach showed that DL based strategis can produce accurate skull-stripping on MPRage and MP2Rage images with *gold* or *silver standard* training data.

7. Acknowledgments

The authors would like to thank Lee Reid and Jason Wood for their significant help and collaboration in this work. Yuliia Kamkova was supported by the Erasmus Mundus Joint Master Degree in Medical Imaging and Applications and by the Commonwealth Scientific and Research Organisation of Australia.

References

Ashburner, J., Barnes, G., Chen, C., Daunizeau, J., Flandin, G., Friston, K., Kiebel, S., Kilner, J., Litvak, V., Moran, R., et al., 2012.

Spm8 manual. wellcome trust centre for neuroimaging institute of neurology, ucl.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26, 839–851.

Avants, B.B., Tustison, N., Song, G., 2009. Advanced normalization tools (ants). Insight j 2, 1–35.

Bernal, J., Kushibar, K., Cabezas, M., Valverde, S., Oliver, A., Lladó, X., 2018. Quantitative analysis of patch-based fully convolutional neural networks for tissue segmentation on brain magnetic resonance imaging. arXiv preprint arXiv:1801.06457 .

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis: I. segmentation and surface reconstruction. Neuroimage 9, 179–194.

Dellaert, F., 2002. The expectation maximization algorithm. Technical Report. Georgia Institute of Technology.

Dolz, J., Desrosiers, C., Ayed, I.B., 2017. 3d fully convolutional networks for subcortical segmentation in mri: A large-scale study. NeuroImage .

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

Galdamesa, F.J., Jailletc, F., Pereza, C.A., 2011. An accurate skull stripping method based on simplex meshes and histogram analysis in magnetic resonance images. Rapport de recherche RRLIRIS 19.

Hahn, H.K., Peitgen, H.O., 2000. The skull stripping problem in mri solved by a single 3d watershed transform, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 134–143.

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.

Iglesias, J.E., Liu, C.Y., Thompson, P.M., Tu, Z., 2011. Robust brain extraction across datasets and comparison with publicly available methods. IEEE transactions on medical imaging 30, 1617–1634.

Ioffe, S., 2017. Batch renormalization: Towards reducing minibatch dependence in batch-normalized models, in: Advances in Neural Information Processing Systems, pp. 1942–1950.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 .

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.

Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M., 2012. Fsl. Neuroimage 62, 782–790.

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Medical image analysis 5, 143–156.

Kalavathi, P., Prasath, V.S., 2016. Methods on skull stripping of mri head scan imagesa review. Journal of digital imaging 29, 365–379.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Kleesiek, J., Urban, G., Hubert, A., Schwarz, D., Maier-Hein, K., Bendszus, M., Biller, A., 2016. Deep mri brain extraction: a 3d convolutional neural network for skull stripping. NeuroImage 129, 460–469.

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.

Lucena, O., Souza, R., Rittner, L., Frayne, R., Lotufo, R., 2018. Convolutional neural networks for skull-stripping in brain mr imaging using consensus-based silver standard masks. arXiv preprint arXiv:1804.04988 .

Marques, J.P., Kober, T., Krueger, G., van der Zwaag, W., Van de Moortele, P.F.ç., Gruetter, R., 2010. Mp2rage, a self bias-field corrected sequence for improved segmentation and t1-mapping at high field. Neuroimage 49, 1271–1281.

Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 3D Vision (3DV), 2016 Fourth International Conference on, IEEE. pp. 565–571.

Van de Moortele, P.F., Auerbach, E.J., Olman, C., Yacoub, E., Uğurbil, K., Moeller, S., 2009. T1 weighted brain images at 7 tesla unbiased for proton density, t2 contrast and rf coil receive b1 sensitivity with simultaneous vessel visualization. Neuroimage 46, 432–446.

Mugler, J.P., Brookeman, J.R., 1990. Three-dimensional magnetization-prepared rapid gradient-echo imaging (3d mp rage). Magnetic Resonance in Medicine 15, 152–157.

OBrien, K., Krueger, G., Lazeyras, F., Gruetter, R., Roche, A., 2013. A simple method to denoise mp2rage, in: Proceedings of the 21th scientific meeting, International Society for Magnetic Resonance in Medicine, Salt Lake City.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Ségonne, F., Dale, A.M., Busa, E., Glessner, M., Salat, D., Hahn, H.K., Fischl, B., 2004. A hybrid approach to the skull stripping problem in mri. Neuroimage 22, 1060–1075.

Smith, S.M., 2002. Fast robust automated brain extraction. Human brain mapping 17, 143–155.

Tudorascu, D.L., Karim, H.T., Maronge, J.M., Alhilali, L., Fakhran, S., Aizenstein, H.J., Muschelli, J., Crainiceanu, C.M., 2016. Reproducibility and bias in healthy brain segmentation: comparison of two popular neuroimaging platforms. Frontiers in neuroscience 10, 503.

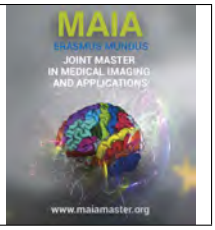
Viviani, R., Pracht, E.D., Brenner, D., Beschoner, P., Stingl, J.C., Støcker, T., 2017. Multimodal memprage, flair, and r2* segmentation to resolve dura and vessels from cortical gray matter. Frontiers in neuroscience 11, 258.

Wachinger, C., Reuter, M., Klein, T., 2017. Deepnat: Deep convolutional neural network for segmenting neuroanatomy. NeuroImage .

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.

Medical Imaging and Applications

Master Thesis, June 2018



Limb Movement Prediction using Subthalamic Nucleus Local Field Potentials for Neuro-Prosthetics

Saed Khawaldeh[a,b,c], Gerd Tinkhauser[d,e], Syed Ahmar Shah[d], Peter Brown[d]

[a]*Erasmus+ Joint Master Program in Medical Imaging and Applications, University of Girona, 17004 Girona, Spain*
[b]*Erasmus+ Joint Master Program in Medical Imaging and Applications, UNICLAM, 03043 Cassino FR, Italy*
[c]*Erasmus+ Joint Master Program in Medical Imaging and Applications, University of Burgundy, 21000 Dijon, France*
[d]*MRC Brain Network Dynamics Unit, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford OX3 9DU, U.K.*
[e]*Department of Neurology, Bern University Hospital and University of Bern, 3010 Bern, Switzerland*

Abstract

As opposed to conventional deep brain stimulation (DBS), directional DBS offers higher spatial resolution and better possibility of more targeted stimulation and decoding. This, consequently, enables us to get more information content from subthalamic nucleus (STN) which could be useful in brain computer interface applications (i.e. communication for those in locked in syndrome) or brain machine interface systems (i.e. neuro-prosthetics for patients with paralysis). Directional deep brain local field potentials (LFPs), also, would offer a greater ability to resolve different limbs from decoding STN LFPs. This study presents a machine learning (ML) approach to investigate whether directional LFPs recorded from STN have sufficient spatial resolution to allow the differentiation of upper and lower limbs which can be used to control neural prosthesis. The results reported in this work prove the efficiency of applying ML on directional LFPs to disentangle upper limb from lower limb during movement preparation and execution. This increases the potential of using LFPs for end-effector selection in neuro-prosthetics.

Keywords: Deep Brain Stimulation, Brain Computer Interface, Neural Decoding, Neuro-Prosthetics, Subthalamus, Nucleus, Machine Learning

1. Introduction

Majority of Brain machine interface (BMI) and brain computer interface (BCI) systems extract brain signals, process, and transform them into commands to control peripherals responsible of carrying out specific tasks, or to communicate with computers which can perform certain processes. One of the major aims of BCI and BMI systems is to replace or restore helpful functions to people with disabilities, especially for those with neuro-muscular disorders such as amyotrophic lateral sclerosis (ALS), stroke, or spinal cord injury. Starting from basic demos of Electroencephalography (EEG) based spellers and single-neuron-based tool control (Birbaumer et al., 2015; Cecotti, 2010; Hochberg et al., 2006), the usage of EEG, intracortical, Electrocorticography (ECoG), and local field potentials (LFPs) started increasing for more advanced tasks including control of robotic arms, prostheses, wheelchairs, and other assistive technologies (Lazarou et al., 2018). BCI and BMI systems can additionally be used for rehabilitation after various brain disorders (i.e. stroke) (Cecotti, 2010).
These systems have two main types, one is using invasive methods and the other is using non-invasive ones. Non-invasive systems primarily take advantage of EEG to control machines (Hinterberger et al., 2005; Kübler et al., 2001). For example, a tetraplegic patient used BCI system which uses beta waves acquired from sensorimotor cortex to grip objects using upper limb robot prosthetics (Birbaumer, 2006). In another example, imagery-based system proposed by Pfurtscheller et al. (2003) was coupled to an implanted neuro-prosthetics hand designed by Kübler et al. (2005) to partially asset patient with paralysis. On the other hand, invasive BCI and BMI techniques are using mainly a group of single brain cells or the neural activity of multiple neurons, this is based on the new electrophysiological methodologies development in chronic, multi-site,

multi-electrode recordings (Keith et al., 1989; Kübler et al., 2005; Nicolelis et al., 1995).
These systems' applications are rapidly growing, however, this growth relies crucially on the availability of convenient, safe, and flexible signal-acquisition hardware. The commonly used noninvasive acquisition hardware (i.e. EEG), and even invasive ones (i.e. ECoG) have certain limitations. For example, EEG is not convenient nor flexible, also it can't be used at all times (i.e. when patient is sleeping or exercising). On the other hand, ECoG is not relatively safe nor convenient, and using it considered to be risky since a large skull opening should be performed to enable placing the grid of electrodes (Shih et al., 2012). Therefore, a better recording technique which can satisfy the essential requirements and provide easy, flexible, and stable neural signal-acquisition is needed. Hence, deep brain LFPs recording using multi-contact electrodes, it can be a robust signal-acquisition technique which avoids the other techniques' limitations and provides potentially a better spatial resolution.
In this paper, we aim to study whether LFPs recorded from subthalamic nucleus (STN) can be used to distinguish upper limb from lower limb at three different periods; at resting before the audio cue where patient is asked to perform a certain limb movement, after the cue and before the movement onset, and during the limb actual movement. Furthermore, we investigate whether the audio cue and the actual movement can modulate the STN in a way that could be used to tell that an audio cue was given or a movement was occurred. Finally, we explore how the important features for classification develops from period to period, for disentangling upper limb from lower limb, pre-cue from post-cue, and pre-movement onset from post-movement onset. The ability to answer these questions contributes toward developing an efficient neuro-prosthetics for paralyzed people using directional deep brain LFPs and ML approach. Also, it discovers what other information, apart from movement related, is encoded in the human's STN.

2. State of the art

Significant development and contributions in neuro-prosthetics has been made by researchers and companies around the globe due to the high potential for these technologies in reinstatement of motor functions in paralyzed patients (Nicolelis et al., 1997; Nicolelis and Ribeiro, 2002). Directional LFPs recording offers a promising option for neuro-prosthetic applications where STN neural signals can be utilized for controlling robot limbs for patients with paralysis or limb loss. Additionally, It is safer and less risky than ECoG where a limited burr hole is needed to insert the electrode for LFPs recording. Also, as it is implanted, it is more stable and more robust than EEG (Giannicola et al., 2012).
STN deep brain recording allows acquiring signals for various types of movements, as STN has representations of different types of movements in one small place, unlike in cortex where movement representations are distributed at larger area (Nicolelis, 2001). Therefore, directional LFPs recording offers a promising option for neuro-prosthetic applications where STN neural signals can be utilized for controlling robot limbs for patients with paralysis or limb loss.
Tan et al. (2016) presented a system capable of decoding gripping force using STN LFPs. Features representing power in the two frequency bands; 55-90 Hz and 13-30 Hz, were the most important ones in decoding the gripping force. These features were fed to first order dynamic linear model to decode the force. However, the force prediction did not work in about half of the study's patients, due to the limitations related to disease impairments, post-operative stun effects, and failure to record LFPs from the 'motor' STN.
Golshan et al. (2018) presented their work about the classification of five behavioral tasks; speech, finger movement, mouth movement, arm movement, and random segments, using STN LFPs recorded while subjects were performing the tasks. The methodology followed to achieve the goal was using cascaded classifiers in tree-like structure. The classification accuracy reported was ranging between 0.607 and 0.706 using support victor machine (SVM) based classifiers.
Mamun et al. (2015) presented a new approach for identifying movement from STN LFPs. This proposed approach takes advantage of features based on causality related to inter-hemispheric connectivity, and a weighted sequential feature selection methodology, which is designed for datasets with small amount of trials and high variability. The results reported illustrate the high efficiency (average accuracy of 0.815) that the method achieves in disentangling left upper limb from right upper limb. The methodology used in our project for disentangling upper limb from lower limb is inspired by the methodology in this work, however, our approach aims at exploring which other information, apart of movement one, can be decoded from STN LFPs to increase the potential of using deep brain recording for neuro-prosthetics.
The work proposed in this project aims to classify lower limb from upper limb, not only during the movement execution, but also during the movement preparation when there is no muscle activity. This would show that STN does not encode information only while the movement is occurring, but also when the subject receives audio cue or prepares for movement. Thereby, LFPs recorded during periods other than movement one, like when person is preparing for movement or intending to move, could hopefully be used to classify upper limb from lower limb, which can be used for end-effector selection to increase the potential use of deep brain recording for neuro-prosthetics application. En-

coding limb from LFPs' movement preparation period is not reported in literature since researchers usually focus on the time around the movement onset.

3. Material and methods

3.1. Experimental paradigm and data recordings

In this study, 5 patients (8 hemispheres) with Parkinson's disease (PD) undergoing STN-DBS surgery were implanted with Boston-Vercise directional lead (Boston Scientific, Marlborough, MA). LFPs were measured intra-operatively after placing the lead in its final position in the STN. First LFPs were recorded with the patient at rest for about 120 seconds. Then the patients have been instructed to perform a block of upper and lower limb movement, as part the routine intra-operative assessment. Within each block the patient performed 15-20 movements of contralateral upper and lower limb. Electromyography (EMG) and accelerometer sensors were mounted on the limbs to record and reliably recognize the single movement episodes for correct labeling.
Figure 1 shows the LFP, EMGs, and accelerometers signals which were recorded while performing the intraoperative tasks of upper and lower limb simple movement. To the lower right corner of the figure, the used deep brain multi-contact lead, and to the upper right corner of the same figure, a simple illustration on the recording and stimulating setup, and where the electrodes are inserted for the LFPs measurement.
Spike2 software (CED, Cambridge, UK) was used to manually label the appearance of the audio cue, movement onset, and end of movement based on the EMG and accelerometer signal. MATLAB (2017b, Mathworks, Natick, MA) was used for segmenting trials, pre-processing and further analyses. Figure 2 shows how the time windows for the upper/lower limb classification were extracted. Figure 3 shows how the time windows for pre-cue/post-cue and pre-movement/post-movement classification within upper and lower limb tasks blocks were extracted.

3.2. Pre-processing

After labeling of the data, they were imported into MATLAB, and detrended by subtracting mean to remove any trend of systematic increase or decrease. The eight LFP channels were subsequently convolved with a complex Morlet wavelet (Cohen, 2014), which aims at transforming the time-domain LFP signal into a time-frequency signal.
This step is needed to calculate the frequency-based features which are the power values in various frequency bands over time. In the wavelet transform, a linear frequency scale of 500 frequency points ranging from 1 Hz to 500 Hz and a variable number of cycles as a function of frequency were used. Finally, four seconds length trials were segmented in time-domain and frequency-domain for each class.

3.3. Feature extraction

A set of 200 ms long non-overlapping windows were used to extract frequency-domain and time-domain features from the eight LFP channels. The power in thirteen different frequency bands ranging from 1 Hz to 500 Hz were identified as potential frequency domain features. These distinct frequency bands are: 1-4 Hz, 5-7 Hz, 8-12 Hz, 13-20 Hz, 21-30 Hz, 31-45 Hz, 46-55 Hz, 56-95 Hz, 96-105 Hz, 106-200 Hz, 201-300 Hz, 301-349 Hz, and 350-500 Hz. Before the extraction of power in these bands, a baseline normalization was performed on each trial following the approach by Shah et al. (2016), where each frequency component in each trial in the time-frequency domain was normalized by a baseline of 750 ms taken from the beginning of each trial before the cue or the movement onset. The baseline allows expressing any STN neural changes presented in the LFPs as a percentage change with respect to the pre-cue or the pre-movement onset LFP, It works by canceling any effect of background activity and allowing comparison across various hemispheres and patients.
On the other hand, regarding the time-domain features, three of them were selected according to the approach in (Hjorth, 1970), where few metrics characterizing the amplitude/time pattern of EEG were presented. These features capture statistical properties of each time-window LFPs, which are not captured by the afore-mentioned frequency-domain features. These metrics were: activity which is a variance measure for the signal; equivalent to the total power in the frequency domain, mobility which is a standard deviation measure for the slope of the original signal relative to the standard deviation of the signal; equivalent to the standard deviation of the power spectrum along the frequency axis, and complexity which is a smoothness measure of the signal with reference to the 'softest' signal that can be computed using the standard deviation of the second derivative (Hjorth, 1970).
Additional to these three features, other seven statistical time-domain features were extracted. These features included mean, standard deviation, skewness, kurtosis, maximum value, minimum value, and entropy. Similar to the frequency-based features, time-based features were calculated for every time window independently.

3.4. Feature selection

After extracting all twenty three features, they were standardized because they vary highly, and it was necessary to normalize them to ensure good classification performance. This features' standardization was applied through making the values of each feature having a zero-mean; by subtracting the mean in the numerator,

and a unit-variance; by dividing by the standard deviation (Grus, 2015).
In other words, the distribution mean (M) and the standard deviation (SD) for each feature were first calculated, then each feature (X) was subtracted by M, and finally the values calculated of each feature was divided by SD as shown in the equation below:

$$X_{standardized} = \frac{X - M}{SD} \tag{1}$$

k-fold cross validation is a common strategy to make sure that all trials in the original training dataset are used for both training and validation; each trial is used for validation only one time. It is used in our case to assess the implemented machine learning model in better way especially because we had limited numbers of data samples. Where k refers to the number of groups which the dataset is to be split into. In this work, we had four folds, where at each rotation, three folds were used for training, and the remaining one was used for testing.
A feature selection algorithm was used to determine the important features and eliminate the non-important ones. This algorithm is called *ReliefF*, it was adopted from (Robnik-Šikonja and Kononenko, 2003), and it was applied on the training subset which contains three folds, then the selected features were used for testing on the fourth fold.

ReliefF algorithm calculates the weights for features through penalizing the ones that give different values to neighbors of the same class, and rewards ones that give different values to neighbors of different classes.
ReliefF sets first all features weights W_j to 0. Then, it iteratively chooses a random trial x_r, finds the k-nearest trials to x_r for each of the two classes, and updates weights, for each nearest neighbor x_q, all the weights for the features F_j as follows:
If x_r and x_q belong to same class,

$$W_j^i = W_j^{i-1} - \frac{\Delta_j(x_r, x_q)}{m} \cdot d_{rq} \tag{2}$$

If x_r and x_q belong to different classes,

$$W_j^i = W_j^{i-1} + \frac{P_{y_q}}{1 - P_{y_r}} \cdot \frac{\Delta_j(x_r, x_q)}{m} \cdot d_{rq} \tag{3}$$

Where W_j^i is the weight of feature F_j at the ith iteration. P_{y_r} is prior probability of the class to which x_r belongs, and P_{y_q} is prior probability of the class to which x_q belongs. m is the number of iterations. And d_{rq} is the distance function of the form.

3.5. Decoding algorithm

In this work, we used a Naive Bayes classification to differentiate between two classes in set of classification problems including upper/lower limb classification, pre-cue/post-cue classification, and pre-onset/post-onset classification. In typical two class classification problems, one class is labelled as 0, and the other is labelled as 1. In this algorithm, each feature or class combination is a separate and independent multinomial random variable (Cohen, 2014).
The estimated probability is given by the equation below:

$$P(C_j \backslash F_1, F_2, F_3, ..., F_d) = P(F_1, F_2, F_3, ..., F_d \backslash C_j) \cdot P(C_j) \tag{4}$$

Where the estimated probability, can be called the posterior probability, which is equal to the prior probability $P(C_j)$ multiplied by the likelihood which is calculated based on the values of the features $F_1, F_2, F_3, ..., F_d$.

This classification method was used because it is simple, and it converges quicker than discriminative models such as logistic regression. Also less amount of training data is required for Naive Bayes, which matches the properties of our dataset especially regrading the low number of trials. The Naive Bayes classifier relies heavily on the conditional independence assumption, where if it holds, it performs very well, and if it does not, it still often does a good job in practice.

3.6. Validation

To be certain about model generalization, 4-fold cross validation was used as explained earlier. The optimal model with the optimal set of features was chosen during the cross validation while applying the ReliefF algorithm on the three training folds. To be sure that the model is not over-fitting, three procedures were performed to validate the obtained classification results.
First, for 20 iterations, labels of the actual segmented trials from the two classes were shuffled, then model was trained and tested, and average across all iterations was calculated. Second, random signals with Gaussian noise resembling the LFP channels were generated, then the whole process of extracting and selecting features followed by testing and training for two class problem, was performed. Finally, random labels of the two classes were assigned to resembled trials extracted from the rest period which lies at the beginning of each subject's recording before performing any tasks, then the whole process was applied again.
In all these three steps, the area under the curve (AUC) values of all the channels were almost equal to the AUC

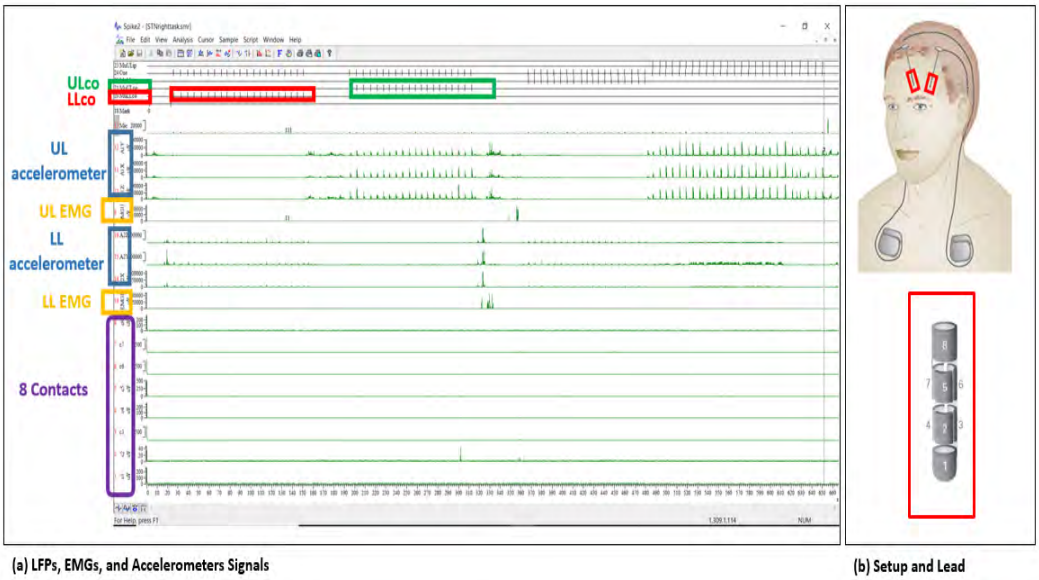


Figure 1: **(a) The LFP, EMGs, and accelerometers signals recorded while performing intraoperative upper limb contralateral (ULco) and lower limb contralateral (LLco) movement tasks, and (b) illustration on electrodes inserted in human's brain to make the recording and the Boston-Vercise directional LFP lead**

value of the random classifier which is 0.5. These validation steps confirmed that the results obtained during the actual classification of the LFP channels are because of the recorded neural activity from the STN, and not because of any sort of noise or over-fitting.

4. Results

4.1. Upper/Lower limb classification

Figure 4 shows the average AUC values across all hemispheres and all LFP channels for upper/lower limb classification for four different periods; rest before the blocks of upper and lower movement tasks, rest within the blocks of tasks before cue, between cue and movement onset, and between movement onset and movement stop. Each point in the figure corresponds to a time window belongs to one of the four periods mentioned above.
Figure 5 and Figure 6 show the AUC values of all the LFP channels, in two hemispheres, for upper/lower limb classification for the four periods explained above. Each set of vertical points in these figures correspond to a time window belongs to one of these four periods.

Figure 7 shows the box plot of the average absolute values of the EMG amplitude across all the subjects and all the trials for upper and lower EMG channels for the three periods; rest before the block of tasks, rest within the block of tasks before cue, and between movement onset and movement stop.
Figure 8 and Figure 9 show the box plots of the absolute values of the EMG amplitude across all the trials, in two subjects, for upper and lower EMG channels for the three periods; rest before the block of tasks, rest within the block of tasks before cue, and between movement onset and movement stop.

Figure 10 shows the average spectrum across all the subjects and all the trials for the down-sampled upper and lower EMG channels for the three periods; rest before the block of tasks, rest within the block of tasks before cue, and between movement onset and movement stop.
Figure 11 and Figure 12 show the spectrum across all the trials, in two subjects, for the down-sampled upper and lower EMG channels for the three periods; rest before the block of tasks, rest within the block of tasks before cue, and between movement onset and movement stop.

4.2. Pre-cue/Post-cue classification

Figure 13 show the average AUC values across all hemispheres and all LFP channels for pre-cue/post-cue classification within upper and lower limb blocks. The twenty three bars under each point in the figure correspond to the weights -based on occurrence- for the features used during the classification.
Figure 14 and Figure 15 show the AUC values of all LFP channels for pre-cue/post-cue classification within upper and lower limb blocks for two different hemispheres. The twenty three bars under each set of vertical points in the figures correspond to the weights -based on occurrence- for the features used during the classification.

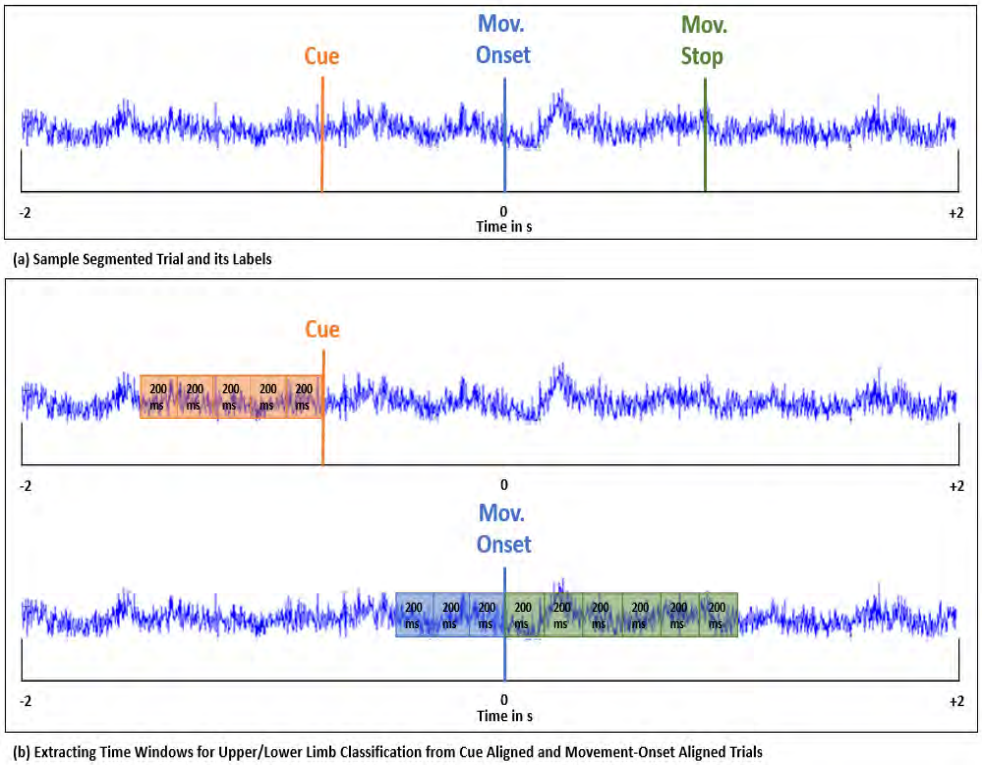


Figure 2: **(a) Segmented and labeled trial sample, and (b) the procedure of extracting time windows from cue aligned and movement-onset aligned trial to perform upper/lower limb classification**

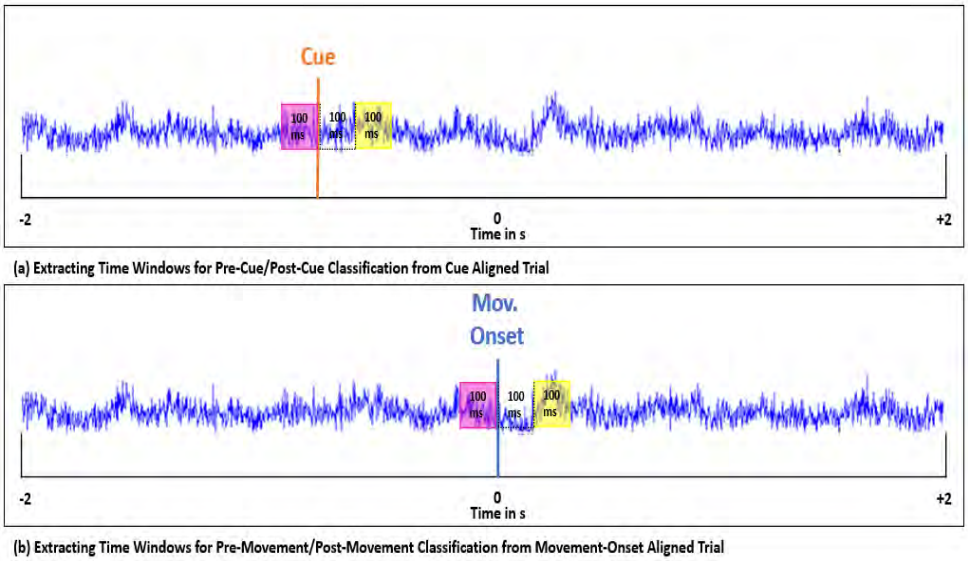


Figure 3: **(a) the procedure of extracting time windows from cue aligned trial to perform pre-cue/post-cue classification, and (b) the procedure of extracting time windows from movement-onset aligned trial to perform pre-onset/post-onset classification**

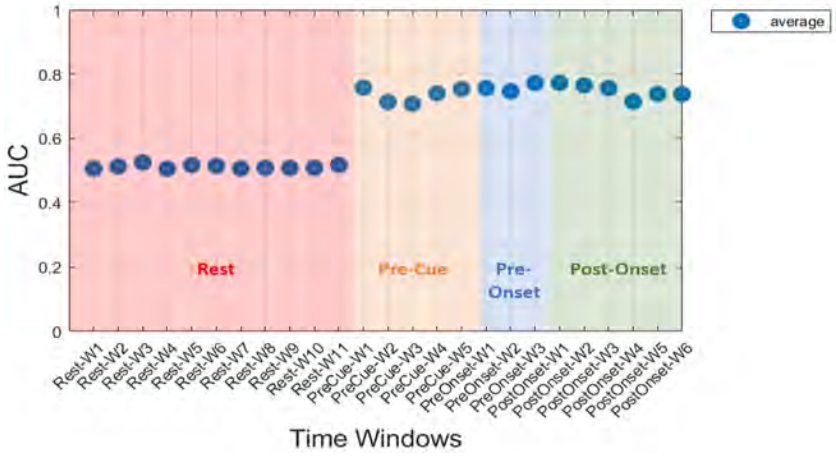


Figure 4: **Average AUC values averaged across all subjects and channels for upper/lower limb classification in four periods; rest, pre-cue, pre-onset, and post-onset**

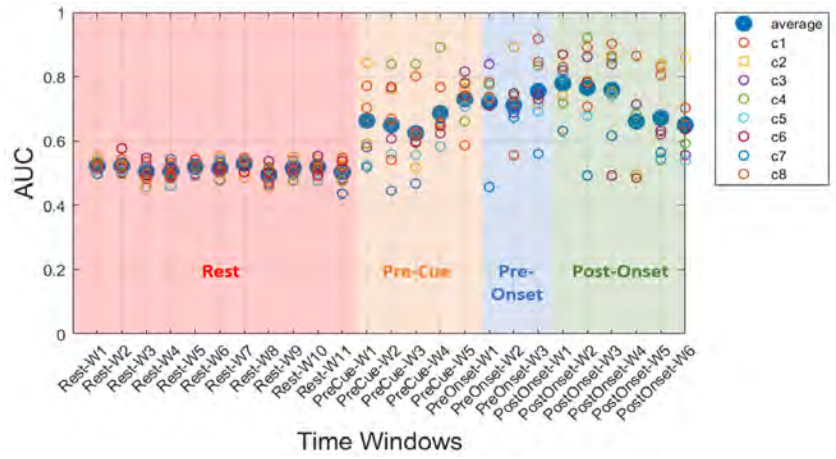


Figure 5: **Average AUC and individual LFP channels AUC values of hemisphere 6 for upper/lower limb classification in four periods; rest, pre-cue, pre-onset, and post-onset**

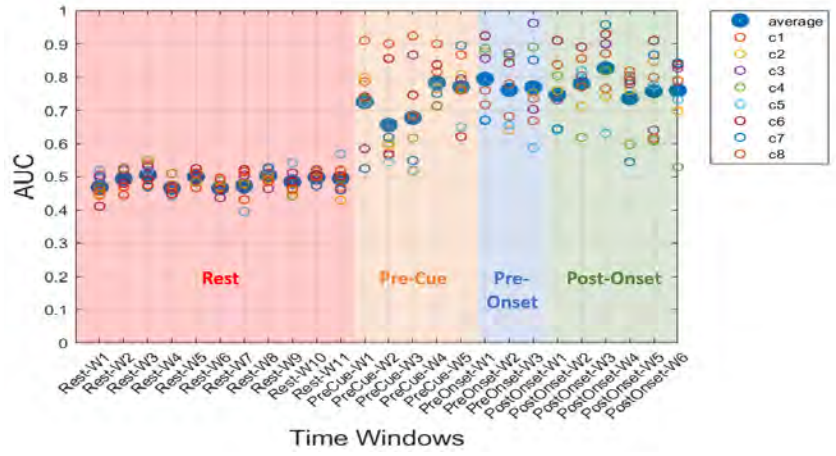


Figure 6: **Average AUC and individual LFP channels AUC values of hemisphere 7 for upper/lower limb classification in four periods; rest, pre-cue, pre-onset, and post-onset**

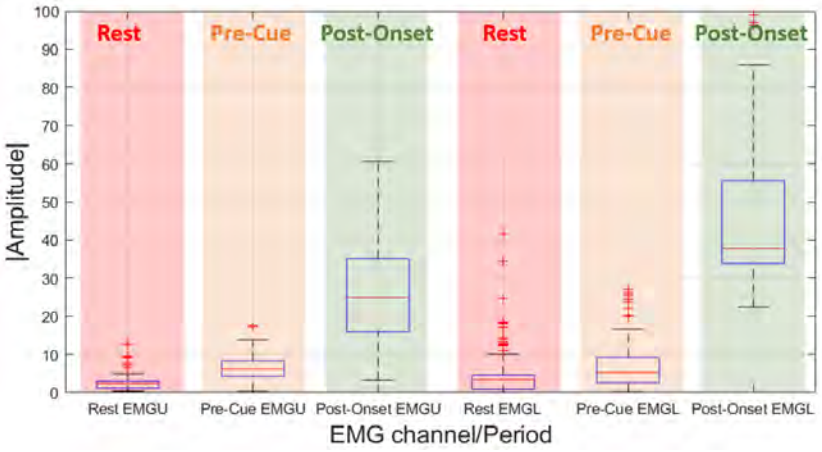


Figure 7: **Average of absolute amplitude for EMG channels across all subjects and all trials before and during upper and lower limb tasks in three periods; rest, pre-cue, and post-onset [U: upper limb, and L: lower limb]**

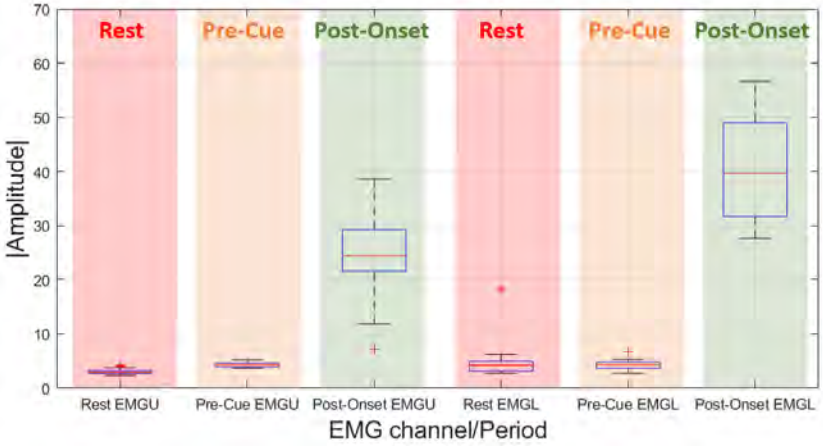


Figure 8: **Average of absolute amplitude of EMG channels across all trials in case 6 during upper and lower limb tasks in three periods; rest, pre-cue, and post-onset [U: upper limb, and L: lower limb]**

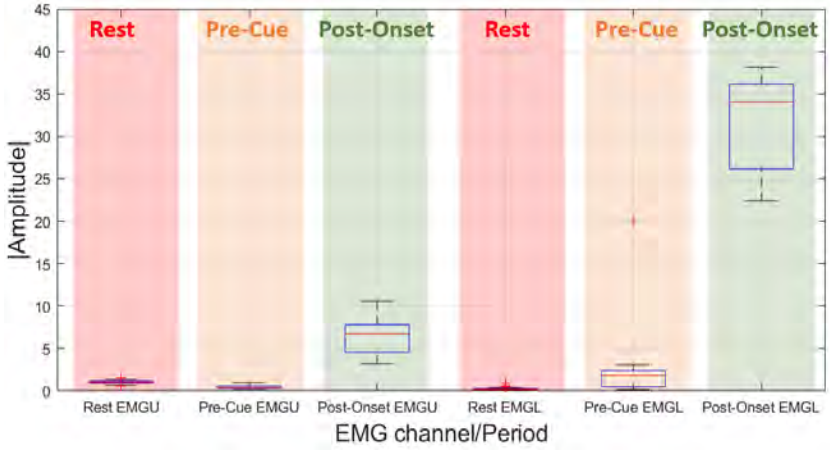


Figure 9: **Average of absolute amplitude of EMG channels across all trials in case 7 during upper and lower limb tasks in three periods; rest, pre-cue, and post-onset [U: upper limb, and L: lower limb]**

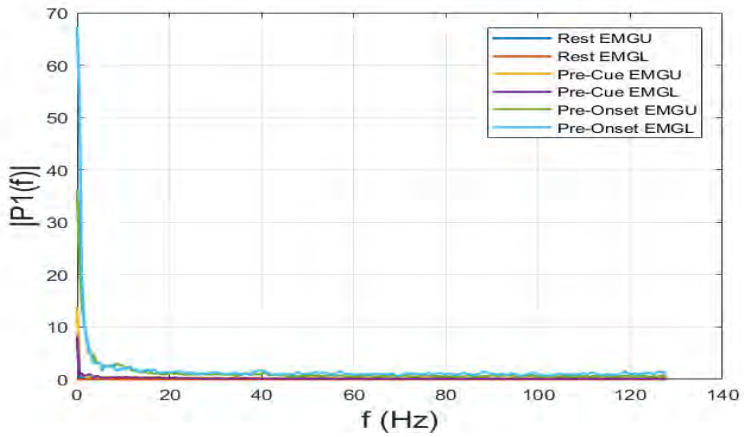


Figure 10: **Average spectrum of EMG channels across all subjects and all trials before and during upper and lower limb tasks in three periods; rest, pre-cue, and post-onset [U: upper limb, and L: lower limb]**

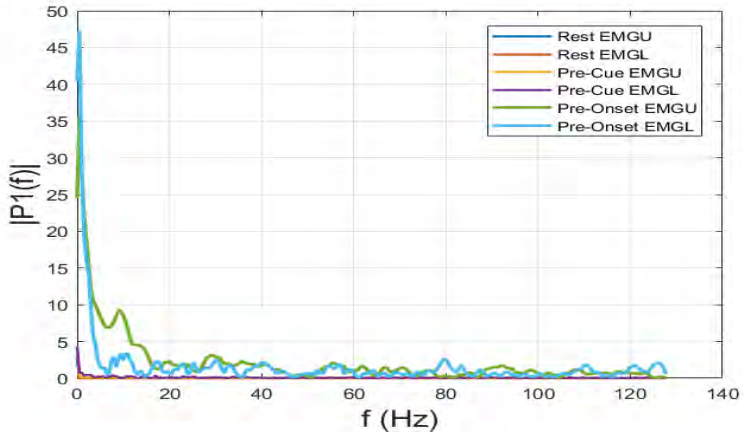


Figure 11: **Spectrum of EMG channels across all trials in case 6 before and during upper and lower limb tasks in three periods; rest, pre-cue, and post-onset [U: upper limb, and L: lower limb]**

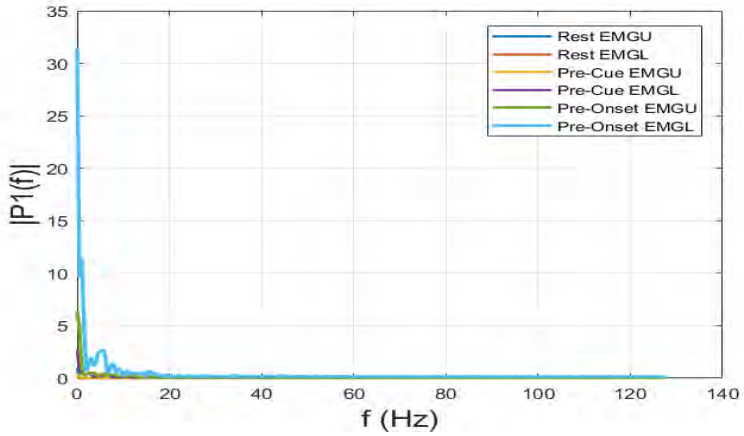


Figure 12: **Spectrum of EMG channels across all trials in case 7 before and during upper and lower limb tasks in three periods; rest, pre-cue, and post-onset [U: upper limb, and L: lower limb]**

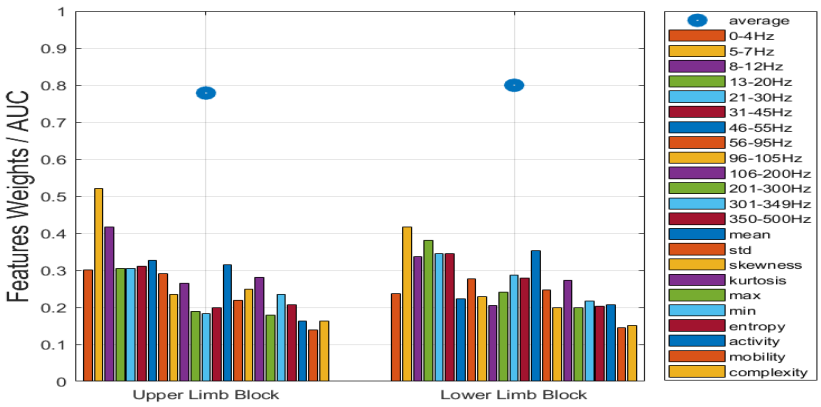


Figure 13: **Average AUC values and average features' weights across all hemispheres and all LFP channels for pre-cue/post-cue classification within upper and lower limb blocks**

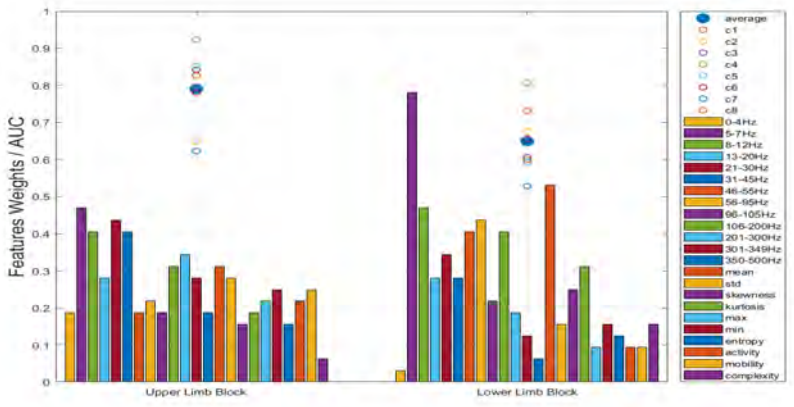


Figure 14: **AUC values and average features' weights across all LFP channels in case 6 for pre-cue/post-cue classification within upper and lower limb blocks**

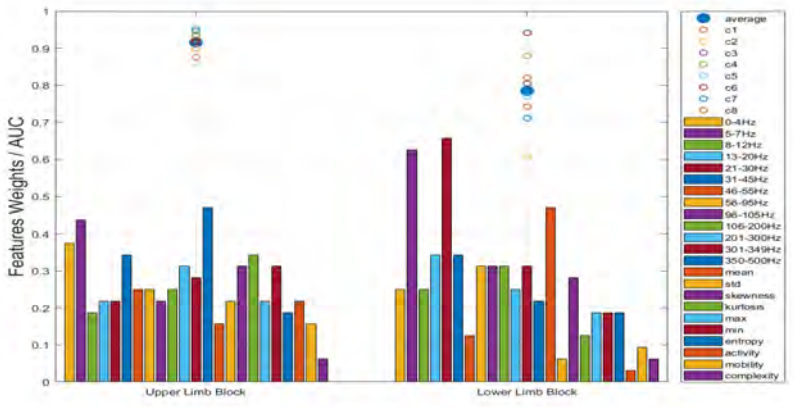


Figure 15: **AUC values and average features' weights across all LFP channels in case 7 for pre-cue/post-cue classification within upper and lower limb blocks**

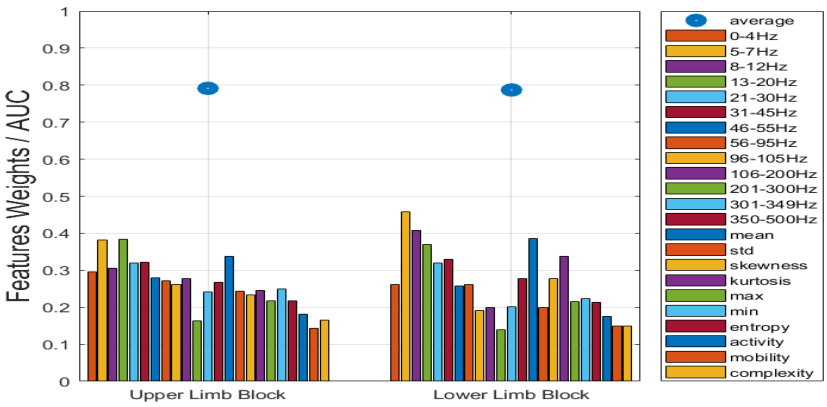


Figure 16: **Average AUC values and average features' weights across all hemispheres and all LFP channels for pre-onset/post-onset classification within upper and lower limb blocks**

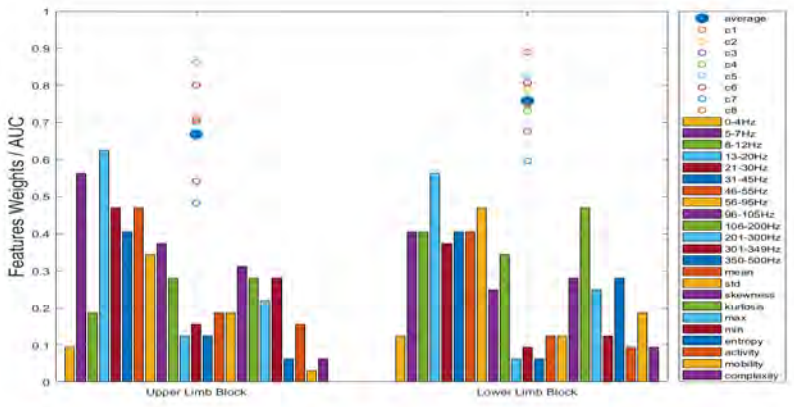


Figure 17: **AUC values and average features' weights across all LFP channels in case 6 for pre-onset/post-onset classification within upper and lower limb blocks**

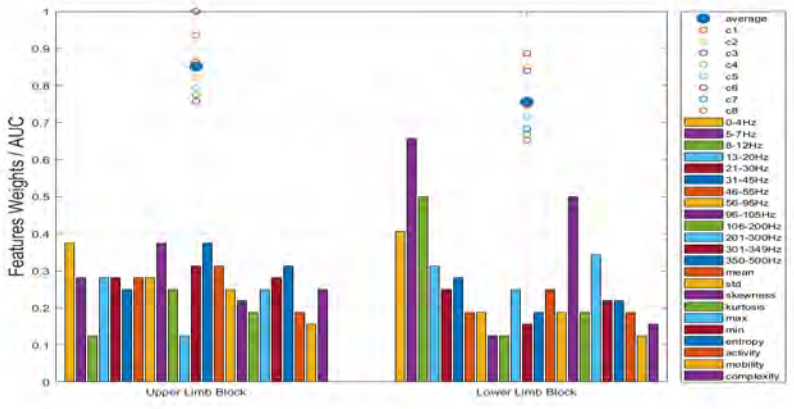


Figure 18: **AUC values and average features' weights across all LFP channels in case 7 for pre-onset/post-onset classification within upper and lower limb blocks**

4.3. Pre-onset/Post-onset classification

Figure 16 show the average AUC values across all hemispheres and all LFP channels for pre-onset/post-onset classification within upper and lower limb blocks. The twenty three bars under each point in the figure correspond to the weights -based on occurrence- for the features used during the classification.
Figure 17 and Figure 18 show the AUC values of all LFP channels for pre-onset/post-onset classification within upper and lower limb blocks for two different hemispheres. The twenty three bars under each set of vertical points in the figures correspond to the weights -based on occurrence- for the features used during the classification.

4.4. Features development

Figure 19 show the average weights across all hemispheres and all LFP channels -based on occurrence- for the features used during the upper/lower limb classification during the three periods; rest before cue, between cue and movement onset, and between movement onset and movement stop.

5. Discussion

In this work, we showed that LFPs recorded from STN encode information about the moved limb, not only during the actual movement, but also during the rest before the audio cue, and also between the cue and the movement onset. This implies that human's STN is neuromodulated differently according to the type of movement which the person performs, and it shows also that muscle activity is not the only reason for the neural information encoded in the STN modulation, to the contrary, the subject's pre-set, preparation for movement, and maybe intention to move the limb, do encode information in the STN.
This change in neural activity in STN allows ML methodologies (i.e. Naive Bayes) to differentiate efficiently between upper and lower limb movements. These different types of STN neuromodulation, enabled us to perform upper/lower limb classification during three periods; pre-cue, pre-onset, and post-onset which suggests that limb prediction can be performed whether there is a movement of not.
As a control, to validate the finding, the upper/lower limb classification task was performed on eleven time-windows while the subjects were at resting before performing any tasks. Results showed that the classification AUC values for all LFPs contacts were random (AUC is around 0.5), since subject's STNs have not encoded yet any movement, preparation to move, or intention of movement. This finding proves also that the algorithm used is not over-fitting.
To quantitatively and qualitatively prove that the ability to perform upper/lower limb classification in the periods other than where the actual movement takes place, was because of STN neuromodulation, not because of any muscle activity, EMG channels absolute amplitudes were calculated and plotted for rest period before any movement, for rest within task blocks before the cue, and during the actual movement between movement onset and movement stop.
Results in Figure 7, Figure 8, and Figure 9 show that there is no significant difference in muscle activity between the two resting periods, however, the difference is significant if the two resting periods are compared to the period of the movement. This finding supports our claim which states that the high performance of upper/lower limb classification, in pre-cue and pre-onset periods, is because of neural information stored in STN, and not because of a muscle activity. As another confirmation, the spectrum of the down-sampled EMG channels during the three periods mentioned earlier were plotted as it appears in Figure 10, Figure 11, and Figure 12. Again, the spectrum plots show a significant difference between the spectrum of the movement period and the two resting periods' ones, which supports the claim regarding the ability to distinguish upper limb from lower limb at the three periods.
These results give more aspiration for the usage of LFPs to control prosthetics for paralyzed patients or people with limb loss, especially because they show that having a muscle activity or performing an actual movement are not the only way to predict upper from lower limb since patients might not have any muscle activity, thereby system needs to rely fully on movement intention or movement preparation to select the end effector for the robot arm or leg which can highly increase functions that paralyzed patients can do.
Another interesting finding appeared in the results of upper/lower limb classification is the change in the channels that gives the highest AUC values across the three different periods; pre-cue, pre-onset, and post-onset. For example, in Figure 5, channels c2, c4, and c6, channels c3, c2, and c1, and channels c6, c4, and c1, provide the highest AUC values in the three periods, pre-cue, pre-onset, and post-onset, respectively. This observation shows a change in the LFP channels which gives the highest AUC values for upper/lower limb classification task. Most importantly, it illustrates the spatial distribution of the eight LFP channels in the recording electrode, and it suggests having more than one neural circuit in STN involved in encoding upper and lower limb movement preparation and execution.
Second and third sections in results show whether there is a discriminating difference in the STN LFPs before and after the cue, and before and after the movement onset. Figure 13, Figure 14, and Figure 15 show that our used machine learning algorithm can efficiently distinguish when the person has received an audio cue, and he started preparing to move. Figure 16, Figure 17, and

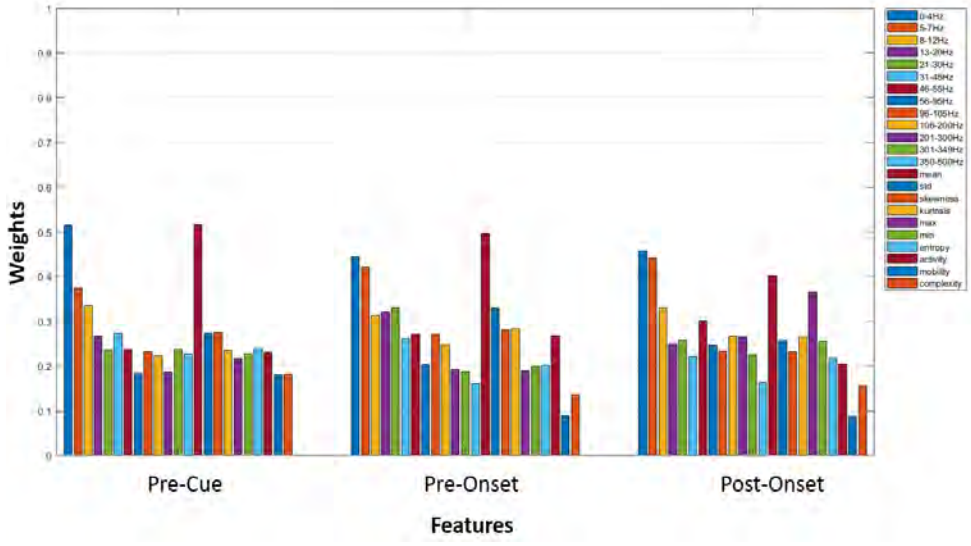


Figure 19: **Averaged features' weights across all hemispheres and all LFP channels for upper/lower limb classification in three periods; pre-cue, pre-onset, and post-onset**

Figure 18, present that we can also distinguish when the person has started moving a limb or not, so it can classify whether the LFP segment is before the movement onset or after it.

These can be further enhanced to predict the expected time of movement before it occurs which can be beneficial for prosthetic application as it would increase the speed and accuracy of the device activation. This would bring us closer to the real-time neuro-prosthetics since predicting the movement before it occurs would contribute in reducing the latency of such systems.

STN neural signals also change when the subject receives an audio cue, and when he moves a limb. This change allowed the implemented ML algorithm to predict cue and movement onset. The prediction works well because of a set of important features which offers the maximum differentiation between the two classes. The most important features are those which offers a maximum separation between the two classes. In this work, the most important six features are selected for testing at each fold, also, they are collected across all channels and folds, then normalized, and finally plotted to show their weights -based on occurrence- as shown in Figure 13 and Figure 16.

Mostly low-frequencies are contributing highly in distinguishing pre-cue from post-cue and pre-onset from post-onset. However, in the single hemispheres, in Figure 14, Figure 15, Figure 17, and Figure 18, features' weights are more variant. In most hemispheres, few of the LFP channels have AUC values higher than 0.9, which suggests that we can rely on one channel for classification, however, it is critical to come up with mechanism to select the best channel. The results show that more analysis on channel level with the help of the electrode mapping -based on the MR intraoperative scans- should be performed to study deeper the features development and the spatial distribution using the recorded STN LFPs and ML approach. Depending on the location of each channel in STN's space, different features might be important for classifying upper/lower limb, pre-cue/post-cue, and pre-onset/post-onset.

Figure 19 shows the weights of the important features -based on occurrence- for the time window with the highest average AUC from each period of the three ones; pre-cue, pre-onset, and post-onset. The figure illustrates that the low frequency features have relatively high importance in upper/lower limb classification. Furthermore, in the post-onset period, the statistical maximum value of the LFP time-domain signal appears to be important which suggests that the LFP channels amplitude varies significantly when performing upper and lower limb movements, which means that one has higher amplitude than the other.

It is worth mentioning, that the results reported in this paper do not utilize history information nor combine LFP channels data to improve performance, instead they are generated from single time window and one LFP channel at each time. Despite the simple used methodology, still it provides relatively high performance for the tasks discussed earlier. Following more advanced solution in analyzing the data would definitely increase the performance further and illustrate the high potential in this promising technology for BMI and BCI application.

Invasive recording in healthy patients for research purpose is not permitted, therefore, LFPs was recorded in PD patients as they have already an implanted electrode for stimulation. In this work, we were assuming that PD and paralyzed patients have similar LFPs, however, this might not be the case. Still, the concept of neu-

ral decoding is the same, and if the classification works for PD patients, it would probably work also for people with paralysis. One of the other main challenges encountered during this study was the limitation of data trials for every class as the recordings were performed during clinical routines where only limited trials are performed by the patients, therefore, it was not possible to use advance machine learning algorithms (i.e. neural networks) to solve the proposed classification problem. Another difficulty faced in this work is regarding defining the features which are to be extracted from the time windows. The approach followed was typically based on extracting twenty three time-based and frequency-based features, but, what if there are few other important features which has not been extracted, thereby, using and automated method to extract the features like convolutional neural network or autoencoders would provide a significant help in discovering new important features, but the low number of data samples in the dataset prevents us from pursuing this approach.

6. Conclusions

In conclusion, our study demonstrated the feasibility of using Naive Bayes ML approach to distinguish upper limb from lower limb using STN LFPs recorded by directional deep brain multi-contacts electrodes. Furthermore, we could predict the audio cue and the movement onset within the one block of certain limb movement tasks. The spatial distribution of channels across the three periods suggests that there are two or more neural circuits in the STN involved in the encoding of movement preparation and execution. The successful prediction reported in this work shows the high potential of using the deep brain recording for neuro-prosthetics which will increase the life standards for patients with physical disabilities.
In term of future work, performing experiments with randomized set of upper and lower limb movement tasks is essential to investigate the prediction of limb movement. Additionally, as the results in this work presents the ability to predict limb while subject is in resting, it is important also to perform experiments of imaginary upper and lower limb movement, and see if we still can predict limb. Finally, performing feature quantification in temporal domain to study the most important features in each channel and how their weights are developing from period to period would be important extension of this work.

7. Acknowledgments

This project was done in collaboration with University of Bern (Switzerland) where data was recorded. Authors thank the Department of Neurology and Neurosurgery of the University Hospital Bern for the acquisition of the original data. S.K. would like to acknowledge the funding by Erasmus Medical Imaging and Applications Joint Program (Education, Audiovisual and Culture Executive Agency Grant 2016).

References

Birbaumer, N., 2006. Brain–computer-interface research: coming of age.

Birbaumer, N., Piccione, F., Chaudhary, U., 2015. Brain-computer interface (bci) communication in the locked-in: A tool for differential diagnosis, in: Assessing Pain and Communication in Disorders of Consciousness. Routledge, pp. 116–139.

Cecotti, H., 2010. A self-paced and calibration-less ssvep-based brain–computer interface speller. IEEE Transactions on Neural Systems and Rehabilitation Engineering 18 number=2,, 127–133.

Cohen, M.X., 2014. Analyzing neural time series data: theory and practice. MIT press.

Giannicola, G., Rosa, M., Servello, D., Menghetti, C., Carrabba, G., Pacchetti, C., Zangaglia, R., Cogiamanian, F., Scelzo, E., Marceglia, S., et al., 2012. Subthalamic local field potentials after seven-year deep brain stimulation in parkinson's disease. Experimental neurology 237, 312–317.

Golshan, H.M., Hebb, A.O., Hanrahan, S.J., Nedrud, J., Mahoor, M.H., 2018. A hierarchical structure for human behavior classification using stn local field potentials. Journal of neuroscience methods 293, 254–263.

Grus, J., 2015. Data science from scratch: first principles with python. " O'Reilly Media, Inc.".

Hinterberger, T., Veit, R., Wilhelm, B., Weiskopf, N., Vatine, J.J., Birbaumer, N., 2005. Neuronal mechanisms underlying control of a brain–computer interface. European Journal of Neuroscience 21, 3169–3181.

Hjorth, B., 1970. Eeg analysis based on time domain properties. Electroencephalography and clinical neurophysiology 29, 306–310.

Hochberg, L.R., Serruya, M.D., Friehs, G.M., Mukand, J.A., Saleh, M., Caplan, A.H., Branner, A., Chen, D., Penn, R.D., Donoghue, J.P., 2006. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. Nature 442, 164.

Keith, M.W., Peckham, P.H., Thrope, G.B., Stroh, K.C., Smith, B., Buckett, J.R., Kilgore, K.L., Jatich, J.W., 1989. Implantable functional neuromuscular stimulation in the tetraplegic hand. The Journal of hand surgery 14, 524–530.

Kübler, A., Neumann, N., Kaiser, J., Kotchoubey, B., Hinterberger, T., Birbaumer, N.P., 2001. Brain-computer communication: self-regulation of slow cortical potentials for verbal communication. Archives of physical medicine and rehabilitation 82, 1533–1539.

Kübler, A., Nijboer, F., Mellinger, J., Vaughan, T.M., Pawelzik, H., Schalk, G., McFarland, D.J., Birbaumer, N., Wolpaw, J.R., 2005. Patients with als can use sensorimotor rhythms to operate a brain-computer interface. Neurology 64, 1775–1777.

Lazarou, I., Nikolopoulos, S., Petrantonakis, P.C., Kompatsiaris, I., Tsolaki, M., 2018. Eeg-based brain–computer interfaces for communication and rehabilitation of people with motor impairment: A novel approach of the 21st century. Frontiers in human neuroscience 12, 14.

Mamun, K., Mace, M., Lutman, M., Stein, J., Liu, X., Aziz, T., Vaidyanathan, R., Wang, S., 2015. Movement decoding using neural synchronization and inter-hemispheric connectivity from deep brain local field potentials. Journal of neural engineering 12, 056011.

Nicolelis, M.A., 2001. Actions from thoughts. Nature 409, 403.

Nicolelis, M.A., Baccala, L.A., Lin, R., Chapin, J.K., 1995. Sensorimotor encoding by synchronous neural ensemble activity at multiple levels of the somatosensory system. Science 268, 1353–1358.

Nicolelis, M.A., Ghazanfar, A.A., Faggin, B.M., Votaw, S., Oliveira, L.M., 1997. Reconstructing the engram: simultaneous, multisite, many single neuron recordings. Neuron 18, 529–537.

Nicolelis, M.A., Ribeiro, S., 2002. Multielectrode recordings: the next steps. Current opinion in neurobiology 12, 602–606.

Pfurtscheller, G., Müller, G.R., Pfurtscheller, J., Gerner, H.J., Rupp, R., 2003. âĂŸthoughtâĂŹ–control of functional electrical stimulation to restore hand grasp in a patient with tetraplegia. Neuroscience letters 351, 33–36.

Robnik-Šikonja, M., Kononenko, I., 2003. Theoretical and empirical analysis of relieff and rrelieff. Machine learning 53, 23–69.

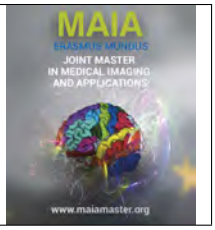
Shah, S.A., Tan, H., Brown, P., 2016. Decoding force from deep brain electrodes in parkinsonian patients, in: Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the, IEEE. pp. 5717–5720.

Shih, J.J., Krusienski, D.J., Wolpaw, J.R., 2012. Brain-computer interfaces in medicine, in: Mayo Clinic Proceedings, Elsevier. pp. 268–279.

Tan, H., Pogosyan, A., Ashkan, K., Green, A.L., Aziz, T., Foltynie, T., Limousin, P., Zrinzo, L., Hariz, M., Brown, P., 2016. Decoding gripping force based on local field potentials recorded from subthalamic nucleus in humans. Elife 5.

Medical Imaging and Applications

Master Thesis, June 2018



Lung nodule classification by means of capsule neural networks

Lev Kolezhuk[a], Keith Goatman[b]

Canon Medical Research Europe Ltd., Edinburgh, United Kingdom

[a] *Leo.Kolezhuk@eu.medical.canon*
[b] *Keith.Goatman@eu.medical.canon*

Abstract

Convolutional neural networks are the core to obtaining solutions with remarkable results to countless computer vision tasks. However, there are some limitations introduced by their concept. One of their main disadvantages is the low preservation of input data in the deeper layers due to simplistic signal routing. Another restraint is their strong tendency to memorize input data, allowing a generically large neural network to express any labeling of the training data while easily fitting random labels. They also suffer from poor generalization, huge training data requirements, low stability to rotational and other geometrical distortions and lack of spatial dependencies and neuron hierarchies. A novel approach of Sabour et al. (2017) has recently been presented that aims to tackle these limitations by replacing conventional max-pooling layers with strided convolutional filters and, moreover, an enhanced hierarchical capsule structure with an advanced dynamic routing algorithm that targets transferring the object perception of the network from global to local coordinate system specific for each object presented in the input. This approach has shown ground breaking results for small handwritten digit recognition, while also having significantly higher generalization to affine transformations of input data than the state-of-the-art CNN.

Our work contributes to the research of this novel concept as well as expands the use of capsule networks to the task of false positive reduction for pulmonary nodule detection in lung CT scans while achieving improved classification accuracy compared to a conventional CNN with max-pooling.

It was also shown that the implemented network is more robust to the change of viewpoints for small images of 3D objects than the baseline CNN. The current architecture of CapsNet has proven not to be easily extendable to data of bigger size and increased overall complexity, while steeply growing amount of trainable parameters and lack of logical constrictions for enforcing the ideological functioning of capsules lead to the vanishing of the conceptual advantages. The comparison of the ability to learn from less training samples was also made for the CapsNet and the baseline CNN with max-pooling. Different configurations of CapsNet were studied and compared in terms of performance in this work. Even though the concept is being at its early development stage, it is able to prove its superiorities and, undoubtedly, further improvements will solve some of the current limitations.

Keywords: MAIA master, Capsule networks, Dynamic routing, Pulmonary nodules, Generalization, Max-pooling, Local perception, Part-whole orientation encoding, Classification, *LUNA16*

1. Introduction

Convolutional neural networks are the current state-of-the-art solution in many computer vision and image analysis tasks. In medical imaging these methods are now beginning to challenge human performance for disease detection and classification. Recent examples include detection of lung disease in chest X-rays from Rajpurkar et al. (2017) and detection of melanoma from Esteva et al. (2017). However, traditional CNN architectures suffer from a number of drawbacks such as

- Poor generalization
- Huge training data requirements

- Low stability to rotational and other geometrical distortions
- Lack of spatial dependencies and neuron hierarchies

We discuss these drawbacks in the next sections.

1.1. Poor generalization

Newer architectures with increased number of parameters are developed relentlessly in order to improve the accuracy of solving a huge variety of upcoming tasks. Nevertheless, growing the depth of neural networks leads to facing the fact that the number of trainable weights becomes many times larger than the quantity of samples in the datasets, by means of which the networks are trained.

Thus, recent studies prove that most recently created deep neural network architectures such as AlexNet and Inception have a strong tendency to memorization and, in fact, do not learn well how one or another object may be represented. Zhang et al. (2016) showed that for the classification task generically large neural networks can express any labeling of the training data while easily fitting random labels and tending to rather memorize inputs that were fed during training than learn the physical function describing valuable features of the input. This study showed that these large networks have enough capacity to memorize each desired output for all of the training images, even if there is absolutely no meaningful correlation between the inputs and their labels in the dataset. Therefore, the networks in the case of fitting random labels were in fact not able to learn how to distinguish one object from another, but build a certain "key-value" relationship database, which allows to route input images to their ground truth labels. Even though, perhaps, it wouldn't be the case for simpler classifiers with less trainable parameters due to their smaller memory capacity, larger and larger networks are developed continuously and the presence of the described problem will eventually lead to a relentless need for increasingly bigger datasets.

Currently the most common and straightforward solution to low generalization is to supply more data for training, while performing augmentation so that the network would learn from as many samples as possible. Unequivocally, obtaining a bigger amount of samples may become a time consuming and quite costly procedure, especially in the regulated areas such as health care, where despite the existence of nation-wide collections of various patient examinations, obtaining access to them is strictly limited by law. Besides, labeling such data requires work of highly qualified specialists that are well familiar with certain medical imaging modalities.

The inability to provide sufficient stability and generalization of AI solutions for disease detection or classification is one of the reasons that the state-of-the-art deep learning approaches are rarely used by clinicians throughout their medical practice. Due to the huge variety of scanners and equipment in hospitals around the world it is one of the greatest challenges to create a solution that would generalize well to this variety of data, while still having a good performance.

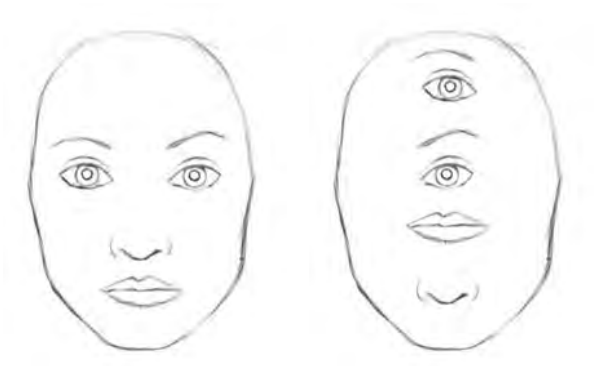
Therefore, it is highly desired to develop a network, that would be able to achieve better performance while requiring less data for learning, and at the same time to offer improved stability to input data changes such as data from a different scanner or another imaging procedure. One of the ways this goal may be fulfilled is by introducing a smarter and ideologically more structured way of learning from data, where each part of the network would perform an operation corresponding to a certain meaning in the physical world such as orientation determination, pose estimation and high level feature detection.

1.2. Low stability to geometrical distortions

Most modern convolutional neural networks provide a certain level of translation invariance. However, frankly speaking, this is achieved by one of their conceptual flaws: loss of information during routing of the signal to deeper layers. Herewith, max-pooling routing is invariant to small translations of input features as soon as the location of the highest signal remains within the same receptive field region, the size of which is defined by the pooling kernel. Despite the fact that this allows step-wise reduction of the feature space dimensionality by removing minor signals with each max-pooling layer, it also means performing data filtering instead of compression. Besides, the translations of the input are not encoded in the outputs of the network, introducing a very little ability of the system to consider spatial relationships between features for making predictions.

If one considers a task of face detection and after training a network on normal non-distorted data, feeds a distorted image(Fig[1b]) for prediction, a conventional CNN due to translation invariance would detect the very same set of features as it would on a non-distorted image((Fig[1a]), generating the same prediction for both images. Because of being invariant but not equivariant to translations, the network doesn't take into account the spatial relationships between features and thus may produce false output as in the given example.

In the domain of medical imaging, sufficiently big neural networks are able to memorize all possible pose combinations of features, that are required to be present in order to trigger certain activation functions and further be used for classification, segmentation or other tasks. For example, for brain MRI clustering into gray matter, white matter and CSF, the network would learn all



(a) Initial non-distorted image of a face. Classified as a face by a conventional CNN

(b) Image of a face distorted with translation of separate parts relative to each other. Also classified as a face by a conventional CNN

Figure 1: Illustration of the disadvantage of translational invariance provided by conventional CNNs with max-pooling: the second image, even though completely distorted, would produce the same output in the network as the image (a) for the face detection task. Since the CNN doesn't consider spatial relationship between eyes, nose and the mouth, both images would trigger the same output because all of the features required for an image to be classified as a face are present.

the possible intensity patterns, that represent edges between every two of the components under all possible orientations. Due to being generically large, the network would be able to perform quite well despite the fact that it considers each orientation of those intensity patterns to be a separate phenomenon, while it is only a certain transformation of a single intensity configuration. Hence, the system would have to learn much more data in this configuration compared to a system that would predict the presence and the pose (transformation, orientation) separately for each feature(Fig[2]).

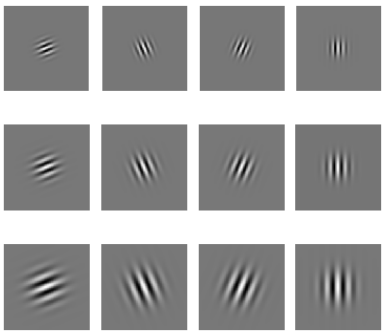


Figure 2: Each of these intensity patterns would be considered to be a separate phenomenon, while it is only a certain set of transformations of a single intensity configuration. Consequently, a conventional CNN would have to learn and memorize all of its possible spatial representations rather that learn it's distinctive feature in the local coordinate system and predict its pose for each of it's representations.

1.3. Huge data requirements and lack of spatial dependencies and hierarchies

It is worth noting that unlike humans, neural networks require hundreds of thousands of samples(if not more) in order to properly learn how to distinguish one object or structural pattern from another. This is partially due to the fact that artificial networks consider the projections of each object on a 2D camera matrix, rather than the objects' distinctive details in respect to one another. Networks that require 3D images as input, have the capability to analyze the entire object, however due to the absence of invariance or equivariance to various transformations(rotation, scaling, translation or other deformations), are able to learn only a single set of features, that are visible and represented under the given pose of that object, but not the general physical properties of that object. This means that a network has to be fed by inputs with all those possible transformations in order to properly learn and afterwards be able to perform well at its task on novel unseen data samples, that might have one of the enormous set of transformations included.

Therefore, there is a necessity in a more sophisticated base for neuron activation mechanisms, that would physically represent not the simple presence of a certain lower level feature, but also take into account how these features are located in respect with each other, hence learning the structure of the objects present on the image in their local coordinate systems.

This thesis investigates a CNN architecture proposed recently that claims to address these drawbacks. Several experiments were performed to examine the performance of such networks.

2. State of the art

For the domain of medical imaging and various disease detection and classification tasks a network that would be able to analyze spatial relationships between lower level features might introduce significant improvements in accuracy, while being able to extract and utilize more complex distribution patterns in order to create and fit higher precision decision hyperplanes. Moreover, the robustness to a variety of transformations may help reducing the demand for bigger datasets, due to the network being able to learn more advanced feature representations from fewer samples. The described features would be valuable for every domain since it would solve some of the greatest problems of the artificial intelligence. These issues are claimed to be tackled by the novel capsule architecture.

Despite the idea of capsules themselves is claimed to have existed for quite a few years in the research group of Geoffrey Hinton, the architecture as well as the novel

dynamic routing algorithm, that allowed such a network to be implemented, have been first presented only a couple of months ago by Sabour et al. (2017). Their experiments have shown significant improvement in classification accuracy and robustness to affine transformations compared to state-of-the-art approaches. Their experiments were mainly performed for *MNIST* handwritten digit classification task, which is a relatively non-complex data. A more recent work, of Hinton et al. (2018) has introduced a different routing algorithm and another structure of capsules. Even though the presented classification accuracy on *smallNORB* dataset in this paper has increased, the performance on the *CIFAR-10* dataset, on the contrary, has become lower. The last introduces a dramatic decrease of the overall network trainable parameter set by approximately 6 times, however to our knowledge this architecture and the claimed performance have not yet been replicated in any other published papers or reports.

In this work properties of the novel architecture such as its robustness to transformations, performance on complex data, features of the training phase, robustness to data distortions, changes of performance with training set size, stability to novel viewpoints and other will be presented. A set of experiments will be performed to explore the functioning of the architecture on complex data. Moreover the task of false positive reduction for lung nodule detection will be targeted by means of this novel system.

Since this architecture is new, not a lot of valuable research has been conducted by now. Therefore, during this work, many discoveries as for the advantages and limitations of this novel network architecture had to be explored before the actual targeting of the final goal.

The architecture implemented in this paper is based on the work of Sabour et al. (2017). The input of the network is a single channel image of size $I \times I$. By means of a 2D convolutional layer low-level features are extracted via A channels with a kernel of size K_1, a stride of $K_{1stride}$ and ReLU activation function. While being fundamentally different, the following layer(Primary Capsules) is essentially a convolutional layer with D channels, a kernel of size K_2 and a stride of $K_{2stride}$. However, the output is separated into D capsule layers, each containing C-dimensional blocks, generated by a defined kernel size K_2 and stride $K_{2stride}$ from the output of the first layer. The size of these blocks corresponds to the output size a convolution generates for a single channel and is the number of capsules in a single layer.

The idea behind the structure of capsules is that each of them encodes the level of certainty that a definite feature is present in the input as well as the pose of this feature in an object-specific local coordinate system. While every capsule in the *PrimaryCaps* layer is being a C

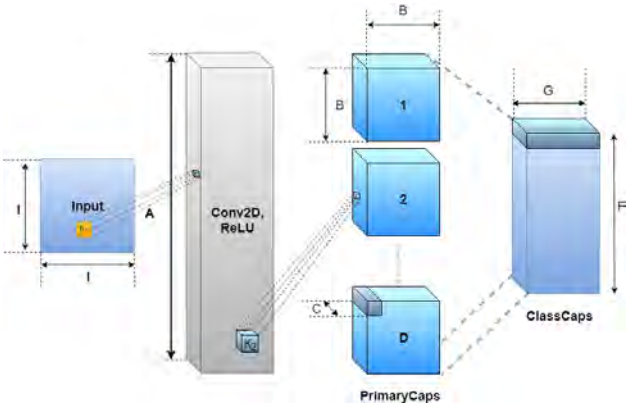


Figure 3: 3-layer capsule network architecture.
$I \times I$ - single channel input image
K_1 - kernel size of the Conv2D layer
A - number of channels in the first Conv2D layer
K_2 - kernel size of the convolution in the Primary Capsule layer
$B \times B \times C$ - size of a single capsule layer
$F \times G$-dimensional digit capsules

dimensional vector, it is desired to encode the probability of feature existence as the length of this vector and a variety of possible spatial transformations in each of its C dimensions. For example, if $C = 8$ a transformation of up to 8 degrees of freedom could be taken into account. Therefore, it would be possible to achieve a certain equivariance to these deformations.

Unlike in a conventional CNN, where the spatial information about features is never taken into consideration, capsules allow to encode and forward this knowledge to deeper layers of the network. It is worth noting that this does not introduce invariance to these specific deformations, but rather equivariance to them, allowing the network to achieve conclusions based not only on presence of certain features, but also on spatial relationship between them.

Figure 4: Illustration of capsule spatial feature encoding for the case of 4 2D capsules. The length of every capsule vector represents the probability of presence of a certain feature in the input signal, while its orientation encodes the pose of this feature in an object-specific local coordinate system.

The goal of every classification network is step-wise dimensionality reduction in each of the consecutive layers, thus it is required to compress the input data size while reaching deeper layers, where higher-level features are aimed to be extracted. Typically the max-pooling operation is performed, that allows filtering the least active elements while forwarding only the most significant ones.

As discussed previously, networks, where max-pooling is chosen for signal downsampling are invariant to generically small translations, but not to transformations with more degrees of freedom.

Alternatively, for the case of encoding both presence and spatial positioning of the features, a novel dynamic routing between capsules algorithm has been introduced by Sabour et al. (2017). It allows the network to make decisions based on agreement between capsules in consecutive layers, due to which the relationship between objects on the input image is dealt with. Every capsule in the *PrimaryCaps* layer is routed to one of the capsules in the *ClassificatonCaps* layer by means of a routing weight matrix $W_{ij} = [C \times G]$. The weights of this matrix are determined by an iterative agreement algorithm.

The *MNIST* model baseline, that was initially used by the inventors of the architecture was configured as given in Tab[1].

Configuration Parameter	Value
I	28
$K_1(K_{1stride})$	9(1)
$K_2(K_{2stride})$	9(2)
A	256
B	6
C	8
D	32
G	16
F	10

Table 1: The original *MNIST* CapsNet model configuration according to the Fig[3] as described in the work of Sabour et al. (2017)

.

The decoder enforces the classification network to learn features in a well structured representation, so that each value of the capsule vector would likely represent a certain physical property of the input. The size of the decoder has to be deep enough to reconstruct most of the input details.

3. Material and methods

3.1. Performance for digit classification

In order to evaluate the performance of our implementation and verify its conceptional correctness, it was decided to perform a variety of tests on the *MNIST* digit dataset. Due to the fact that the mentioned data is definitely not as complex as, for example, medical data, it is unequivocally a good idea to first evaluate the performance of the network for simpler tasks. Since the studied

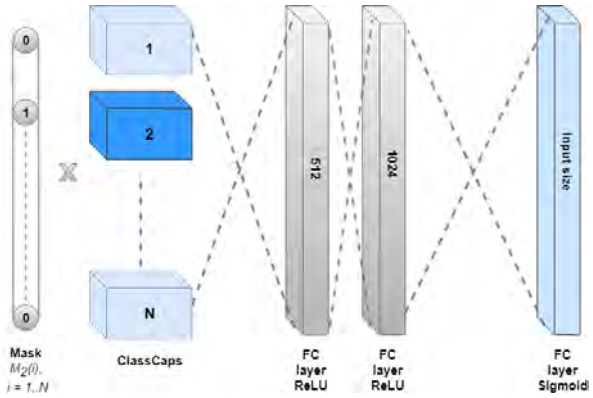


Figure 5: Reconstruction is used to enforce the capsule network to act as an encoder, while learning the most descriptive features of the input. The similarity between the output of the given decoder network and the initial input image is maximized. The *ClassificationCaps* layer is masked by the ground truth labels(during training) or by choosing the longest of all N vectors(during testing) so that the mask is as given

$$M_j(i) = \begin{cases} 1, & \text{if i = j} \\ 0, & \text{otherwise} \end{cases} \text{ where j - target class}$$

The output of the last layer is chosen to be the same as the size of the initial input image $I \times I$.

concept is completely new, many of its properties remain unexposed, the following primary experiments also allowed us to better understand the real potential and current possible applications, where the system would make a noticeable difference compared to the state-of-the-art approaches.

3.1.1. Proof of implementation correctness

It was decided to replicate the results presented in the original paper and also verify the robustness of the network to affine transformations of the data as Sabour et al. (2017) claimed to achieve ground breaking results for this experiment. To explore the robustness of the model to data transformations the model initially trained on the *MNIST* dataset was applied for classification of another sample set from *affNIST*, where random affine transformations were carried out for each original digit. The last set is previously unseen to the network and thus allows to measure its invariance and robustness to the given transformations.

3.1.2. Feature encoding and reconstruction

The decoder network allows to reconstruct images from the capsule vectors in the *ClassificationCaps* layer. This forces the network to learn the most important and descriptive features that are common for all digits and encode them in a compressed representation of a G dimensional capsule vector. The more detailed this compressed information is, the more precisely each input image may be reconstructed. Once the network converges to a state, when it can properly reconstruct the inputs from their

compressed representations, it becomes more probable that the system would appear more robust to unseen data.

If the data fed to the network was rotated around a certain axis, it would thus be highly probable that a specific part of the capsule vector encodes rotation of the object around this axis. If one would desire generating a model that is robust to affine transformations of the input data, it is likely that this would only be achieved once feeding the network with enough data subjected to these transformations.

This finding correlates with the conceptual disadvantage of modern neural networks, suggesting the remaining presence of a significant flaw in perception mechanisms compared to human vision. The core of this issue lays in the amateur ability of neural networks to extrapolate their knowledge to new incoming data, while one of their biggest advantages being the ability to interpolate it. It is the last one that allows networks to accomplish good results on modern tasks once being trained on a sufficient amount of data.

However, even though data augmentation required to tackle this issue is somewhat computationally demanding, a network that benefits more from it while better learning the object properties would still be of great interest. The aim to physically structure the network into groups of neurons, that is targeted in the architecture of CapsNet, might expand the capacity of networks to extrapolate, while enforcing them to encode spatial orientation of features in capsule vectors and dynamically using determined groups of neurons for each specific input(by means of the dynamic routing algorithm). Owing to the structuring of neurons into groups representing specific physical operations, such networks could be able to deal with previously unseen data in a better way. This would introduce the possible opportunity of training the networks on less data, while still attaining state-of-the-art outcomes.

The way that CapsNet encodes handwritten digits into the capsule vectors of *ClassificationCaps* layer has been investigated. In order to detect and verify the fact that each value in these vectors corresponds to a certain physical feature of the digit, we have performed a study in which random noise is added to the remaining masked vector, therefore simulating a change in the encoded representation as if it was performed by the network itself.

We present the illustration of the observed phenomenons in Sec[4.1.2].

3.1.3. Robustness to affine transformations

The given architecture has also been examined in terms of robustness to affine transformations of the input data. In order to conduct this, the network trained on the original *MNIST* dataset was applied on previously unseen data from the *affNIST* dataset, that consists of *MNIST* images distorted with random affine transformations. In the case of CapsNet being more reliable for this task than a conventional CNN with max pooling, one may conclude that its advanced structure and signal routing indeed introduce improvements to the way the given data is learned and analyzed by the network. It is worth mentioning that for such simplistic data, where the image doesn't contain complex background, object shadows, or small descriptive object details, each digit may be considered to be somewhat of an affine transformation of a single digit of this class. Hence, the network is able to learn on data already subjected to a certain degree of affine transformations, therefore adapting to such changes while learning to recognize these distortions.

The results of this experiment are presented in Tab[2].

3.2. Performance for 3D object classification

The performance of the CapsNet on digit classification datasets establishes evidence that this architecture works rather well on undemanding data(the *MNIST* dataset contains images with no background and the digits themselves do not have any complex micro features). Nevertheless, the idea of encoding the objects' features, their spatial orientation and deformations in the capsule vectors may become problematic for cases when the object is not easily distinguishable from the background, dropping shadows, while being subjected to different lighting configurations. This may lead to the network overfitting to the training data, while the capsules wouldn't encode physical properties of objects' features on a certain common scale of all possible inputs but other information such as simple presence of certain intensity patterns in pixel intensities specific for each input, acting as a conventional CNN with no benefits of logical neuron groups structuring.

Therefore, the architecture was tested on a more complex *smallNORB* dataset, containing images of 3D objects of 5 different types taken under various lighting configuration from a set of defined viewpoints(LeCun et al. (2004)). Since this dataset contains extensive metadata that precisely describes under which conditions every shot was taken, such as azimuth, elevation of the camera in respect to the object, lighting intensity, class subtype and others, it allows performing a selective training on a certain subset of data while measuring the networks' response to changes of input configuration.

The input data was normalized to a zero intensity mean and a unit standard deviation.

3.2.1. Reconstruction

To investigate the effect that complex data has on the quality of reconstruction, and also determine the opti-

mal network configuration, different configurations of the network were tested, e.g. changing the number of channels in the primary convolutional layer in the range of (32, 64, 128, 256, 512) as well as the dimensionality (8, 16 values per capsule) and quantity (8, 16, 32, 64) of Primary Capsule types. The dimensions of capsules in the *ClassificationCaps* layer has been set to 16 or 32 according to the dimensions of capsules in the previous layer. In order to extract deeper features prior to passing the signals to the capsule layers, the number of convolutional layers before the *PrimaryCaps* layer has also been manipulated.

The depth of the decoder network has also been changed in order to verify that there is sufficient reconstruction capability for each given configuration of capsules.

The results of the most significant architecture alterations are presented in Sec[4.2.1].

Figure 6: Sets of azimuths used for training (purple) and testing (gray) of the network. The samples for training and testing were taken from two different partitions of the dataset and have no overlap between them.

3.2.2. Robustness to novel viewpoints

In order to test the invariance of the network to changes in viewpoints, i.e. changes in the position from which the objects were photographed, it was decided to set up two networks: (a) a baseline CNN with max-pooling and (b) the CapsNet. Each architecture was optimized for the task, the training parameters adjusted to maximize the invariance of each network.

The CNN baseline network was created to be of similar size to the CapsNet network in terms of the number of trainable parameters. It consists of two consecutive groups of 2D convolutional layers with 64 filters, a single strided kernel of size 5×5, followed by a max-pooling layer with a pool size of 2×2 with valid padding and ReLU activation functions, two fully connected layers of 1024 neurons each with ReLU activation functions, completed by another fully connected layer with size corresponding to the number of possible classification predictions with a softmax activation function.

Because the *smallNORB* dataset is well structured and contains necessary meta-data for each sample, it is possible to perform the following experiment. Each network is trained on a defined set of viewpoints (containing azimuths of 300, 320, 340, 0, 20, 40) that is approximately one third of all training data in size. Afterwards two tests are taken for each network: one on a set of familiar viewpoints (containing the same azimuths as for training) from the test partition of the dataset and another one on a set of novel viewpoints (containing azimuths from 60 to 280) previously never seen by the networks. This allows us to determine how the performance of each network changes once applied on data, where images of objects were taken from different positions than the ones it was trained on.

The given split of the training and test data allows to test the ability of the network to extrapolate the learned knowledge to unseen data, where the images of objects are taken under unseen points of view (Fig[6]).

The results of undertaken experiments for different model configurations are presented in Tab[4].

3.3. Performance for lung nodule classification

During previous tests it was noticed that the current architecture has a significant decrease in object feature encoding ability with the growth of complexity of the data. Therefore, it is crucial to perform sufficient preprocessing and structuring of the particulars.

While the size of the network and the number of trainable parameters grow steeply with the increase of the input size, it is highly desired to keep the input dimensions limited. This would allow also limiting the number of parameters and hence possibly improving generalization, while also keeping the classification accuracy at its maximum value. Thus, it was decided to process 64×64 images as it was determined that this is the minimal dimensions, for which there is no notable improvement in the network performance when using inputs of bigger size.

As determined by Perandini S. (2016) and others there is a certain distribution of pulmonary nodules inside the lobes of the lungs. These distributions are different for each of the nodule categories. For example, in the mentioned study it was determined that the prevalence of adenocarcinomas and other non-carcinoid cancer types is located in the upper lobe, while the prevalence of carcinoid tumors is located in the middle and right lower lobe with a tendency to occur in the central lung parenchyma.

Quinn Colin Meisinger (2011) discovered that 85% of lesions arise within the central airways as endobronchial

masses, which are associated with symptoms such as cough, shortness of breath, wheezing, hemoptysis. Some other reports found malignancies to favor an upper lobe location. For that reason, it is logical to conclude, that the orientation information (the location of the finding) as well as the spatial relationship between structures surrounding the candidate (e.g. airways, bronchioles, proximity of pleura etc.) may be structured into certain complex statistical patterns, determining which would most probably enhance the performance of detection and classification algorithms for this task.

To utilize the presence of such complex patterns at the highest possible level, the classification algorithm is desired to be able to identify them, taking into account not only the presence of definite structures on the input image, but also their spatial relationships relative to the candidate and to each other.

Speaking of which, capsule networks might break the performance of the state-of-the-art approaches for tasks where these relationships between features are of the utmost importance.

3.3.1. Dataset preprocessing

The *LUNA16* dataset was generated from the *LIDC-IDRI* database that consists of diagnostic and lung cancer screening thoracic computed tomography (CT) scans with marked-up annotated lesions. This screening collection contains 1018 cases, provided by seven academic centers and eight medical imaging companies. The *LUNA16* was created as a subset of the *LIDC-IDRI* by selection of scans with a slice thickness smaller than 2.5 mm, which resulted in 888 scans being included.

The *LUNA16* challenge also provided a list of candidate nodule locations in each of the CT scans, which were generated by a semi-automated detection algorithm. Ground truth was provided that stated for each candidate whether it was a true or false nodule detection. Patches were generated centered on each candidate in all the CT scans. These patches were afterwards classified by the means of the chosen neural network.

The CT volumes were rescaled so that the image voxels were isotropic, i.e. each voxel corresponds to a cube with sides of 0.5 mm. This procedure was performed so that the pose and spatial transformation prediction does not have to take into account the anisotropy of the space where a feature is located, and therefore allows it to encode the object related descriptors with less effort and using a more compact representation. Furthermore, it allows the network to equally process patches of the three different views.

In order to extract the maximum fraction of valuable details that are represented in the 3D space of a CT scan, three orthogonal patches are extracted for each candidate, as illustrated in Fig[7]. This allows a deeper and a more robust analysis of each candidate in comparison with the case of extracting patches only in one defined view (i.e. sagittal, axial or coronal planes).

Figure 7: 3 orthogonal slices of size 64×64 extracted for each nodule candidate from CT volumes. The slices are extracted with 5 different angles per each view of each candidate. This results in 125 patches per positive candidate.

During the first analysis of the samples, one may easily notice a major imbalance between the positive and negative classes since the patient with a pulmonary nodule is luckily a less frequent case than an unaffected healthy person. Because the network fits its weights to maximize a certain performance metric, it is crucial to ensure that both sample categories are considered to be of equal importance. To achieve this the positive class (containing pulmonary nodules) is upsampled to have 125 entries per positive candidate. This is done by applying data augmentation, by rotating each of the three orthogonal views by a set of angles $\in \{-20, -10, 0, 10, 20\}$ degrees, resulting in five images per view per candidate.

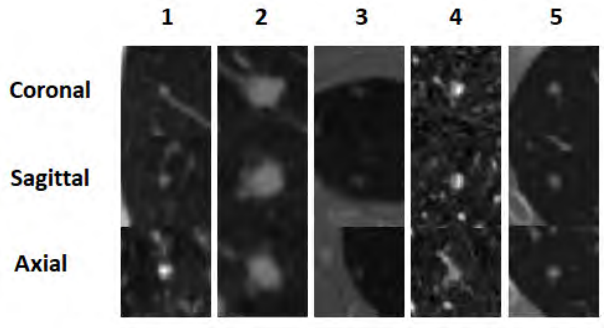


Figure 8: Randomly selected patches in orthogonal planes for 5 different candidates.

The amount of negative candidates, on the contrary, is *decreased* by taking only the first 150 locations for each volume and once again extracting their sagittal, coronal and axial views.

This allows to create a balance between the positive and negative classes distributed closely to 50%:50%, while

generating approximately 750 000 patches. 10% of the data has been dedicated for testing purposes.

When selecting the size of the patches it has to be considered that important information lies not only in the nodules themselves, but also in their nearest surrounding neighborhood. Choosing a sufficient neighborhood size allows the network to base its decisions also on the presence of certain structures, which may indicate regions with a high or low probability of a mass finding, while enhancing the overall accuracy of the predictions.

According to the analysis of the size distribution of pulmonary nodules on lung computed tomography scans performed by Li et al. (2017), most of the malignant nodules have a diameter of 3–6 mm; larger sized nodules occur much less frequently (Fig[9]). Since we have to properly cover the entire scale of diameters, it was decided to choose the patch dimensions so that even the largest nodules would be completely contained within a single patch. Thus according to the classification of lung nodules into miliary nodules (less than 2 mm), pulmonary micronodules (2–7 mm), pulmonary nodules (7–30 mm), pulmonary masses (more than 30 mm), the biggest nodule size to be considered is 30 mm. With a chosen isotropic voxel spacing of 0.5 mm, a 30 mm nodule can be fully contained in a 60×60 patch. While introducing a relatively shallow neighborhood with 64×64 pixel patches, due to the given nodule size distribution, most of them would include larger surroundings.

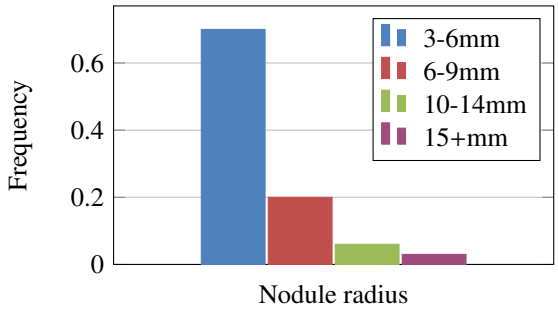


Figure 9: Appearance frequency for different lung nodule diameters in the positive class as given in the study of Li et al. (2017) .

The current state of the art approach to the false positive reduction task on the *LUNA16* challenge uses multiple patches extracted in the same view instead. However this could mean that the network better fits the parameters to data which is more similar, rather than considers the physical meaning of every layout on the image.

3.3.2. Performance with training set size

The CapsNet is an approach that introduces an advanced learning procedure to the conventional neural network. Since one of the goals of artificial intelligence research is to allow networks to learn in a manner closer to that of humans, it is of great interest whether the described architecture is able to distinguish inputs better while being fed less data than a typical network (such as the baseline CNN) requires.

In order to assess this ability to learn from fewer example datasets, an experiment was performed involving testing the performance of the models trained on different numbers of datasets. Each of the models was trained using a certain fraction of the entire training dataset (3.2%, 11%, 33%, 100%), and the performance tested on an unseen set of samples. The experiment was conducted for both architectures (i.e. the CapsNet and the CNN baseline) while the performance was measured by means of calculating the area under receiver operating characteristic curve (AUC).

3.3.3. Training computation time

Due to the CapsNet using a significantly more complex routing algorithm, which in optimal configuration makes three iterations per single data pass, the training time of the CapsNet is significantly higher than for a CNN using max-pooling. All experiments were performed on a server with $2 \times$ Intel Xeon E5645 CPUs, 99 GB RAM, and an NVIDIA Titan V GPU. It was determined that the CapsNet model requires approximately 8.3 times more training time per batch than a conventional CNN with an equivalent number of parameters (Tab[5]).

3.3.4. Robustness to data distortion

Capsule networks have proven to be more robust to adversarial attacks. In a very recent study by Hinton et al. (2018) it has been shown that the capsule network model is significantly less vulnerable to both general and targeted FGSM adversarial attacks (Goodfellow I. (2014)). The model was also tested on the slightly more sophisticated adversarial attack of the Basic Iterative Method (Kurakin A. (2016)), where it was also found that the model was much more robust to the attack than the traditional convolutional model.

Even though in the domain of medical imaging stability to adversarial attacks is not at the forefront of current research, these experiments show that the novel architecture is less sensitive to minor data distortions. This suggests it may also generalise better and be more robust. This is of the utmost importance for healthcare. Therefore, an investigation was made of the robustness of the networks to a variety of plausible data distortions, that may occur in real scenarios of medical imaging.

The primary type of distortions exists due to the fact that the medical imaging data is created by a wide variety of scanners worldwide. Every scanner has its own specific features and software with image processing algorithms embedded within it. Behaviour is model and

manufacturer dependent. This introduces a set of problems: certain details on the images of one scanner may be less visible than on the images of others, entire intensity patterns may change due to different calibration, data compression or contrast enhancement techniques used in each machine.

Thus, in order to develop a framework that could be widely used independently of the scanner manufacturers or the institutions where they are located, one has to either collect enormous sets of training samples in order to feed the network with all possible variations of data during training or, preferably, create a system that is less vulnerable to such changes between hardware and software imaging technologies. For example, in CT scanners one of the main differences among existing manufacturers is the calibration that is used to transform the relative electron densities or stopping powers of tissues to the scale of Hounsfield units (HU). Even though the HU scale is always calibrated relative to the relative stopping power (RSP) of water, which is always set to correspond to be HU = 0, values for other tissues may vary significantly between scanners due to differences in X-ray beam filtration, reconstruction method and post-processing. Cheng et al. (2013) have investigated and compared the HU-RSP calibration curves for 18 CT scanners from Philips, General Electric, Siemens, Toshiba Medical, Picker and Accuray (Fig[10]).

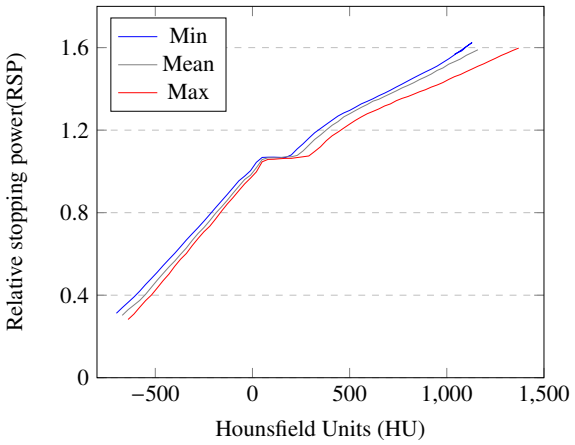


Figure 10: HU-RSP curves for 18 CT scanners: HUmin-RSP, HUmax-RSP and HUmean-RSP, labeled as Min, Max and Mean in the graph. It represents the difference in calibration curves for the tested machines and allows to mimic certain distortions of data in order to test the robustness of the CapsNet to such changes.

According to the mentioned study an intensity mapping function has been created, that introduced the ability to mimic data from different scanners and test the networks' response to such data modification as if images from a new scanner were fed to it during testing. The mapping function used represented a case where calibration curves were distorted to the maximum possible degree according to the difference study. Thus, HU intensities above 0 HU were increased to the maximum possible degree for their range, while intensities below 0 HU were decreased by the corresponding maximum change. Other mapping functions were tested as well, however those cases resulted in a lower network response, so the previously described case may be considered the most extreme possible change of HU-RSP calibration curves.

A secondary type of distortions that causes distinctness between CT samples exists due to the specific CT reconstruction kernels used by each scanner manufacturer. The variance of these reconstruction kernels results in a change in sharpness and noise levels in the volumes following tomographic reconstruction from the raw sinograms. Sharper kernels result in a more significant noise in the target volumes. However this is favoured by some radiologists, as they believe it enhances fine details. Note that each CT scanner has many built in kernels that may be chosen, ranging from sharper to smoother (less noisy) images.

Previously, Gierada et al. (2010) have shown the effects of CT slice thickness and reconstruction kernels on a set of quantitative descriptors of the volumes. Gallardo-Estrella et al. (2016) introduced a technique for normalization of such data, that was created by means of different reconstruction kernels from different scanners. This technique significantly increased the similarity between the volumes of the 369 subjects from the *COPDGene* study in terms of emphysema Scores. Nonetheless, the mentioned normalization technique helps to reduce the variance of a certain quantitative function among the images in the study, but not equalize them in terms of all possible features. This might reduce the effect of the problem, however it does not disappear altogether.

In order to analyze how well the proposed architecture copes with such issues, it was decided to mimic the difference between images by performing sharpening on a subset of the datasets with chosen kernels. Afterwards, both networks that were previously trained on non-distorted data were tested on this generated subset. The sharpening kernels were chosen to be of size 3×3 voxels, with a range of auxiliary coefficients so that

$$K = S_{aux} \cdot K_{sharp}$$

$$S_{aux} = 1, 5, \cdots, 30,$$

where K_{sharp} is the sharpening kernel and S_{aux} is an auxiliary coefficient. Even though this doesn't completely recreate data from different CT machines, it allows to resemble the variance in image sharpness and thus analyze the stability of the networks to it.

3.3.5. Overall performance

In order to generate the most efficient and accurate model, various configurations were tested to optimize the parameters in Tab[1]. The goal is to determine the configura-

tion that provides the best performance, i.e. the most accurate classification of nodule patches in terms of AUC, while minimizing the total number of trainable parameters in the network, to discourage overfitting. Overfitting also unnecessarily increases training and inference run times and reduces the generalization of the model.

Different input dimensions of the data were tested. The initially extracted patches of size 64×64 were downscaled to sizes of 48×48 and 32×32 by means of bilinear interpolation. Lower input sizes offer fewer low-level details than the initial patches. However, the smaller patches vastly reduce the amount of trainable parameters, which may result in better generalization to novel data, while preserving a high classification accuracy.

The influence of the primary convolutional channels, A, that are extracted from the input, has been analyzed. The reduction in this number eases the further processing of the signals by deeper capsule layers and, taking into account the fact that each capsule must represent the presence and the instantiation parameters of a certain combination of lower level features at a given location, increasing the number of these primary channels results in the need to combine more feature types.

The number of primary capsule types, D, influences the potential to process and make predictions of more details and separate groups of features extracted by the first convolutional layer. Theoretically, the more complex and diverse the data is, the more primary capsule types should be used. The real influence of this model parameter on the overall accuracy of the model was also studied.

The dimensions of capsules in the *PrimaryCaps* and *ClassificationCaps* layers correspond to the allowed degrees of transformation freedom, that for each of the objects they may be sufficiently represented by C and G values. If it is required to allow more possible ways of transforming objects of the input, one may consider increasing these dimensionalities. However, the larger they get, the less probable it is that the network will encode the predicted pose, and is highly likely to overfit and act like a conventional convolutional layer. We have studied the real influence of the dimensionality of capsules for various datasets.

We have also created a model that averages the vote of the network for each of the three patches corresponding to the coronal, sagittal and axial planes, in order to predict whether the given candidate location contains a nodule or not. Unlike creating an ensemble of models, one for each of the views, this approach allows us to train a single model on the set of all views and further classify each patch with the same network. The voting procedure averages the prediction weight for each class $v_c(i)$ among all patches for each candidate $i \in 1..3$. Afterwards, the highest averaged prediction weight $max(\tilde{v_c})$ indicates the class to which a candidate is assigned. The prediction for class $c \in (pos, neg)$ is as given:

$$\tilde{v_c} = \sum_{i=1}^{3} v_c(i)$$

The results of these studies, as well as the comparison of the best CapsNet models for the given task, are presented in Sec[4.3.5].

4. Results

The results of the set of performed experiments and studies are given in this section. Each of the subtopics relates to the corresponding entry in the "Materials and methods" section.

4.1. Performance for digit classification

This following experiment was performed in order to verify whether our capsule network behaves similarly to the original network from Sabour et al. (2017).

Applying the network on the *MNIST* digit classification task has proven that the implementation used in the further work manages to perform comparably to the results claimed in the study where the architecture was first described and thus confirms the validity of the further experiments. The test error rates for cases of applying the network on *MNIST* and *affNIST* datasets are presented in the figure Tab[2].

4.1.1. Proof of implementation correctness

Dataset	Test error rate %		
	CapsNet (Sabour et al.)	**Baseline CNN***	**CapsNet (our implementation)**
MNIST	0.25	0.39	0.29
affNIST	21	34	25
multiMNIST	5.2(80% overlap)	5.2(4% overlap)	—

Table 2: Performance of CapsNet on datasets for digit classification. The implementation used in this paper has proved to achieve results, closely comparable to the implementation of Sabour et al. (2017). *Baseline CNN - previous state-of-the-art neural network of Wan et al. (2013). It is worth noting that the mentioned architecture achieved 0.21% test error rate while using an ensemble of 5 networks as well as data augmentation by rotation scaling.

The presented solution was able to achieve very similar performance metrics for the *MNIST* and *affNIST* datasets when compared to the original implementation of Sabour et al. (2017) (Tab[2]). The test error rates achieved with the given implementation were slightly higher that the ones originally presented by the authors, however are still significantly better than the baseline CNN approach.

4.1.2. Feature encoding and reconstruction

Fig[12] shows examples of input reconstructions for some *MNIST* images. In the capsule vectors of the *ClassificationCaps* layer each of the values describes a certain degree of transformation freedom that the object or digit can have, such as stroke thickness, rotation, width, height or stroke type (specific curvatures). Since this last layer of the network is actually masked, either by the ground truth labels in case of training or according to the longest capsule vector of all in case of testing, each of the vectors does not encode the class of object in its values but only the variations of this specific object class. The class itself is described by whichever vector remains non-zero after the masking operation. This gives an extended capability to the encoded structure. The results of randomly changing each of the 16 values of the remaining unmasked vector may be observed in Fig[11].

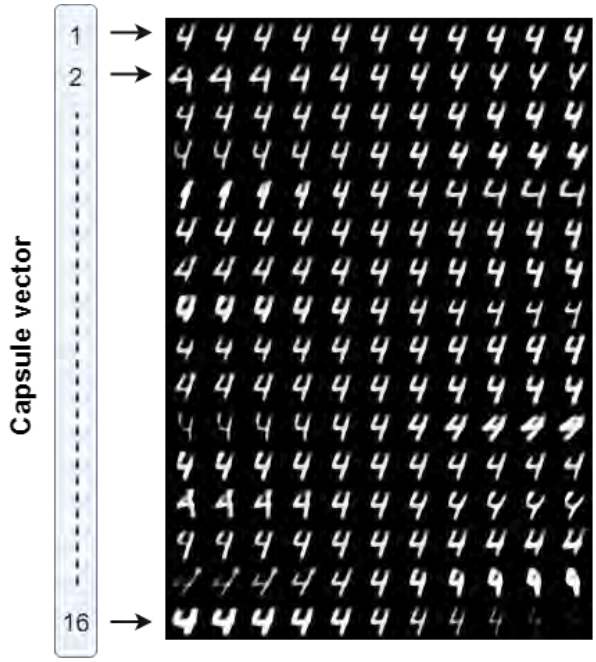


Figure 11: Response of the reconstruction network to random noise introduced to each value of the 16D capsule vector of the *ClassificationCaps* layer in case of digit "4". Even though the mentioned layer is masked, the reconstruction takes into account all of the capsules it contains, however the encoding of the physical properties of a certain digit take place in the single remaining unmasked vector of all. It is this vector that is manipulated.

One may notice that the reconstructed inputs are in fact quite detailed representations of the initial images, meaning that for this data the network allows to properly encode most of the details present on the images in the classification capsules of the last layer. Despite the fact that every value of the capsule vector indeed encodes a certain variation of the specific digit, such as stroke thickness, width, translation and rotation, it is essential to understand that there is still no control over which physical meaning would be encoded in these vectors during training — this relies solely on the distribution of data in the training dataset.

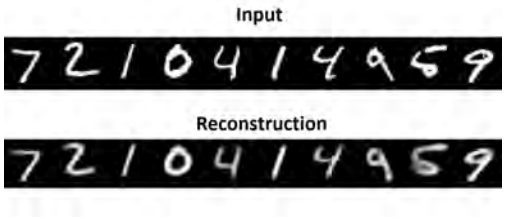


Figure 12: Initial inputs with reconstructions obtained as output of the decoder network, that recreates the image from the information encoded in the capsules of *ClassificationCaps* layer.

4.1.3. Robustness to affine transformations

Once subjected to testing on the *affNIST* dataset with random affine transformations of the data, the accuracy measures obtained showed that the CapsNet has a smaller drop in performance than the baseline CNN model compared to the one on the original *MNIST* set as mentioned in Tab[2].

4.2. Performance for 3D object classification

In this experiment the performance of CapsNet for small images containing 3D objects is studied. The following results show how the capsule network is able to cope with this data of greater complexity compared to the *MNIST*.

Despite the fact that the paper of Sabour et al. (2017) claims that the described architecture achieved 2.7% test error rate on *smallNORB*, our best attempt only resulted in 10.9%. The inability to replicate the results of Sabour et al. (2017) is also seen in the paper of Xi E. (2017), where for the *CIFAR10* dataset instead of 10.6% claimed in Sabour et al. (2017), these authors only achieved a test error rate of 28.5% using an ensemble of four networks and an additional convolutional layer prior to the *PrimaryCaps* (31.07% with the mentioned *MNIST* model baseline). This leads to a conclusion that at this moment of time, there is a gap in performance between the original implementation and others, that might have its cause in advanced data preprocessing techniques or other crucial network configurations, that remain undisclosed.

4.2.1. Reconstruction

If one considers analyzing the ability of the network to reconstruct the input image from the capsule vectors in the last layer of the encoder part, it is clearly visible that for this dataset, the network encodes projection shapes of the objects present on the image, however not much

Dataset	Test error rate %	
	CapsNet	CNN
smallNORB	2.7 (EM Caps - 1.8%)	2.0
	10.9 (our experiment)	
CIFAR-10 (RGB)	10.6 (EM Caps - 11.9%)	4.5
	28.5 (experiment of \cite{capsulesOnComplexData})	

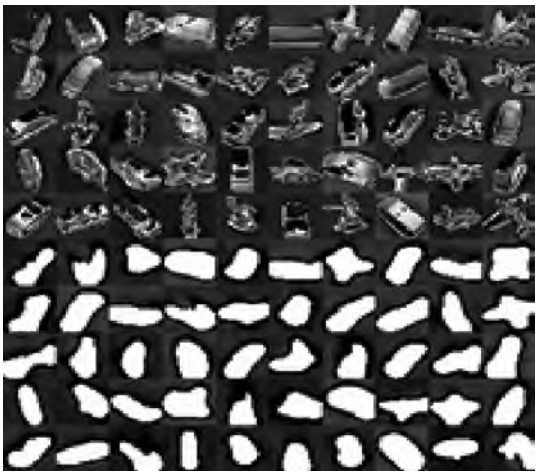
Table 3: Performance of CapsNet and the corresponding state-of-the-art CNN with max-pooling on the *smallNORB* and *CIFAR-10* (RGB images) datasets. Another architecture of Hinton et al. (2018) claims to achieve better accuracy on *smallNORB*, however on the *CIFAR-10* capsules with dynamic routing performed better according to the data from the two papers. The performance on *smallNORB* achieved during our experiment largely differs from the performance originally claimed by Sabour et al. (2017). The inability to replicate the original claim on *CIFAR-10* also appeared in the paper of Xi E. (2017).

details (such as shadows, minor edges) are recreated (Fig[13]).

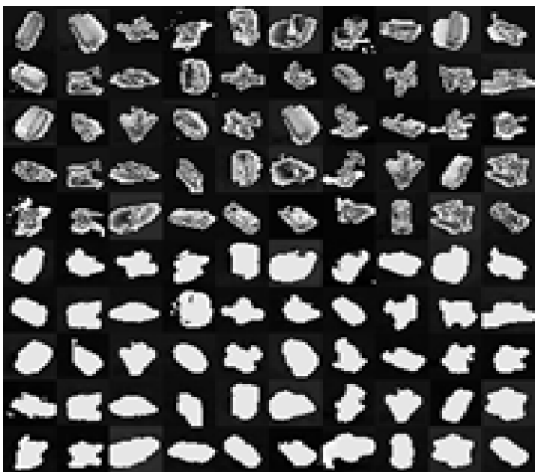
It may also be seen that the larger model with more convolutional channels and more primary capsule types produces better reconstructions (Fig[13b]) than the shallower one (Fig[13a]). However, the decoding did not seem to improve any further while increasing the dimensions of the capsules in the network (from 8 to 16 in the *PrimaryCaps* and from 16 to 32 and more in the *ClassificationCaps* layers). In fact, any attempt to force the network to encode less important features — such as by stacking more convolutional layers in the early stages of the network or changing the dimensions of convolutional kernels, capsule vectors and etc. in order to extract deeper features — didn't improve the encoding and reconstruction noteworthy.

Changing the type of the loss function, that aims to match the reconstructed image with the initial input, did not exhibit any notable changes to the quality of reconstruction.

After many attempts to enhance the classification and encoding performance, the model of Fig[13b] has been chosen for further experiments.



(a) 64 convolutional filters, 8 8D capsules in the *PrimaryCaps* layer



(b) 256 convolutional filters, 32 8D capsules in the *PrimaryCaps* layer

Figure 13: Initial inputs and reconstructed images decoded from the masked *ClassificationCaps* layer capsule vectors for the *smallNORB* dataset. Since the data is more complex the network is not able to encode all of the feature details, but rather shapes of the objects on the inputs.

4.2.2. Robustness to novel viewpoints

A robustness coefficient was chosen as the invariance metric, that is defined as the fraction of test accuracy on a set of novel and unseen viewpoints compared to the test accuracy on a set of familiar ones. The higher the metric, the better the result.

Resulting network performances for both CapsNet and baseline CNN are presented in Table[4]. One may notice that despite CapsNet having 256 convolutional filters in the first layer, and 32 8D Primary Capsule Types, it had a slightly lower accuracy on familiar viewpoints than the best baseline CNN model. However, it has proven to have an impressive performance on novel viewpoints (0.98 against 0.92 in terms of the robustness coefficient).

The architecture of CapsNet provides a classification solution that is thought to be closer to human perception. However, during the experiments above it was noticed data augmentation was crucial in both models, for performance and robustness improvement. While for a conventional baseline CNN with max-pooling this was expected, it was hoped the CapsNet would perform otherwise.

That the capsule network still requires data augmentation may be explained by the need of an actual enforcement for the network to learn in a desired manner. Thus, if a capsule network is trained on data, where all the features have a very limited set of possible orientations and deformations, there is absolutely no technique that would encourage proper pose prediction during the training phase. Therefore, the CapsNet trained in this way is

Network Model		Test accuracy: Familiar viewpoints	Test accuracy: Novel viewpoints	Robustness coefficient
CapsNet 64 conv. filters, 8×8 Primary Capsule Types	32×32 input	82.19	66.78	0.81
	32×32 input augmented	83.04	77.23	0.93
	32×32 input (random crops from 48×48)	84.58	68.44	0.81
	32×32 input (random crops from 48×48) augmented	83.72	79.91	0.95
CapsNet 256 conv. filters, 32×8 Primary Capsule Types	**32×32 input (random crops from 48×48) augmented**	**81.73**	**80.11**	**0.98**
Baseline CNN 32×32 input		80	68.97	0.86
Baseline CNN 32×32 input - augmented		83.96	77.04	0.92

Table 4: Invariance to novel viewpoints for different models of CapsNet and CNN baseline. The robustness coefficient shows the fracture of accuracy on novel viewpoints compared to the accuracy on familiar viewpoints.

never actually learning as it is designed to, while fitting its weights to the training data and using capsules as certain micro internal layers, that learn unpredictable information about the input features, but not their poses.

This discovery brings up an important conceptional limitation of the new architecture, while introducing a need of a training procedure that would encourage actual pose and transformation prediction in capsule vectors rather than abstract information about the features.

Nevertheless, even though data augmentation is still applied during the learning phase, the trained capsule network has a better ability to extrapolate its knowledge to new incoming data than the one of the baseline CNN, resulting in a better robustness to novel viewpoints. Furthermore, data augmentation can always be performed at a low cost, and it's need should not be considered a disadvantage.

4.3. Performance for lung nodule classification

In this section the results of applying the CapsNet for lung nodule candidate classification are presented. We investigate and compare the dependence of performance on training set size for CapsNet and the baseline CNN, study the encoded representation of features and stability of the networks to a set of distortions. The comparison of various configurations of the capsule network is also given.

4.3.1. Performance with training set size

Figure 14 shows plots of nodule classification accuracy (measured by the area under the ROC curve) versus the number of training datasets, for both the baseline CNN and the CapsNet implementations. While the performance of the CapsNet architecture is slightly better for all training set sizes, the behaviour of the two networks is quite similar: a drop in accuracy is almost identical when using less training data and is represented by a pattern common for most of the machine learning systems.

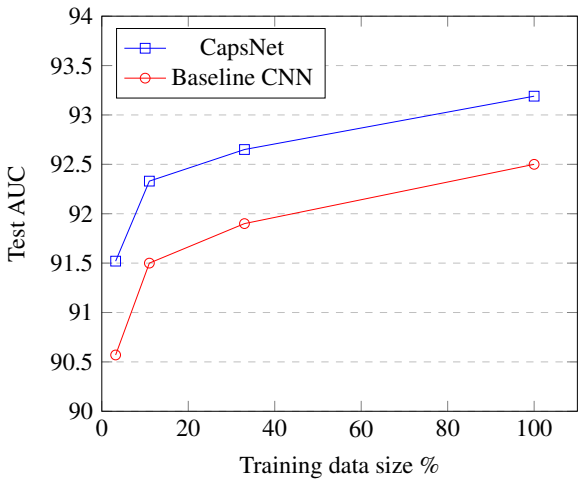


Figure 14: Classification accuracy (area under ROC curve) versus the fraction of the training set used to train the networks. Note that the CapsNet result is superior at all training set sizes.

4.3.2. Feature encoding and reconstruction

The reconstructions of inputs for the case of *LUNA16* lung nodules have proven to contain the most important descriptions of each patch, such as the estimated shape of the nodule candidate, as well as some structures that may surround it (such as the tissue of lung walls, airways, etc.). Unlike the decoded representations for the *small-NORB* dataset, where the level of present details was not sufficient and the reconstructions were able to represent only an estimate of the objects' shape but not their details, the encoding of lung nodule candidate patches preserves the most significant structures.

Figure 15 shows some examples of reconstructions from a CapsNet model with 64 convolutional filters in the first layer, 32 Primary 8D Capsule Types, and ten 16D capsules in the *ClassificationCaps* layer.

4.3.3. Training computation time

While being quite predictable, the finding that the capsule network epoches require more processing time is explained by the complexity of the dynamic routing between capsules — involving time consuming operations with the capsule vectors such as dot product, mean, squashing, soft max operations. At the same time, in the baseline CNN this step is using only a rather simple maximum value selection among each of the given kernels,

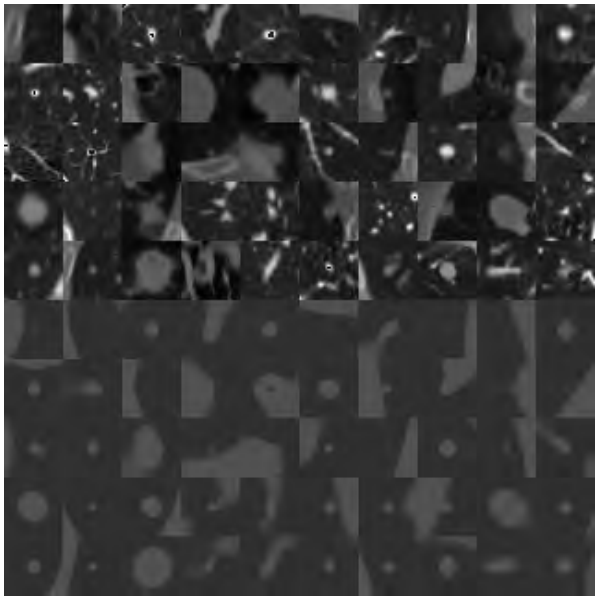


Figure 15: Initial 32×32 inputs (upper half of figure) with the corresponding reconstructions (lower half of figure) for *LUNA16* patches for the CapsNet model with 64 convolutional filters, 32 Primary 8D capsule types. The nodule candidate as well as some of its surrounding structures are visible on the reconstructed images, while smaller details are left not encoded.

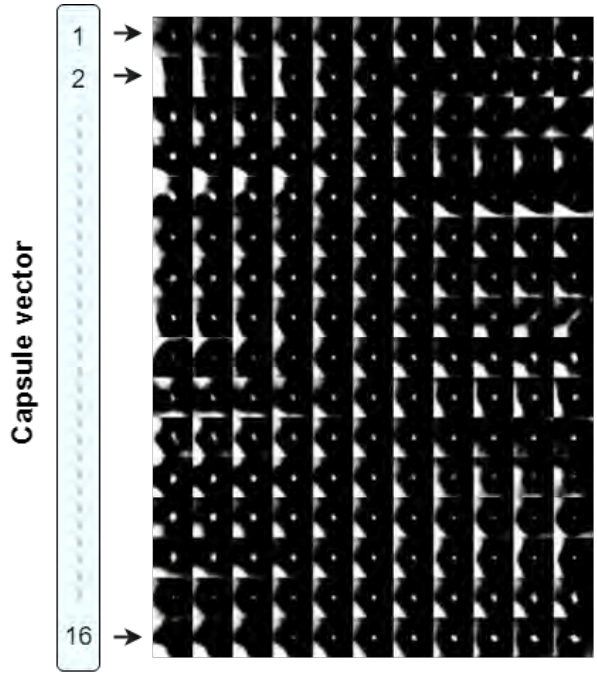


Figure 16: Response of the reconstruction network to random noise introduced to each value of the 16D capsule vector of the *Classification-Caps*. As in the case with the *MNIST*, quite a few values redundantly represent the same physical change to the object on the image, e.g., the size of the nodule itself is encoded in most of the vector values rather than a single value being assigned to this property. Each vector part represents a superposition of a set of physical properties of the input.

Computation time per training step	
CapsNet	**CNN**
158ms	19ms

Table 5: Computation time of one training step(with 32 samples per batch) on *smallNORB* dataset for the CNN baseline and CapsNet architectures. The networks were configured to have a similar number of trainable parameters(~10m). The tests were taken on a machine with 24 Intel Xeon E5645 CPUs, 99 GB RAM and NVIDIA TitanV GPU.

that is relatively undemanding in terms of computation time.

Besides the greater computation time required for each training iteration, the CapsNet also converges slightly less rapidly, requiring more epochs to reach a stabilized classification solution. Therefore, in terms of computation time, CapsNet has a very strong disadvantage once compared with a conventional CNN of a similar number of parameters.

4.3.4. Robustness to data distortion

The response of both CapsNet and the baseline CNN to such distortion appeared to be equal resulting in the performance to drop to 99.9% of the initial AUC on non-distorted data (Tab[6]). Even though the CapsNet did not appear to be more stable in this case, it is worth noting that such a small drop in classification accuracy serves as an indication of the fact that both network types are quite robust to variations in CT intensity calibration.

Distortion type	**Performance change (percentage of initial AUC on non-distorted data)**	
	CapsNet	**CNN**
HU scale change	99.9%	99.9%
Sharpening ($S_{aux} = 5$)	99.88%	99.89%
Sharpening ($S_{aux} = 30$)	95.74%	97.33%

Table 6: Response of the CapsNet and the baseline CNN model to image distortions: HU calibration scale change that mimics data as if it was from another scanner, sharpening of intensities that is performed uniquely in the software of each scanner manufacturer.

According to the results of this experiment both networks have responded to feeding a set of sharpened images with $S_{aux} = 5$ in a similar way (the baseline CNN had a slightly larger AUC decrease compared to the CapsNet). While for a stronger sharpening of $S_{aux} = 30$ the baseline

CNN appeared to be more stable, resulting in a smaller response (Tab[6]).

4.3.5. *Overall performance*

In this section the performance of some of the tested CapsNet models as well as the baseline CNN models are presented. These models correspond to various configurations of CapsNet with different capsule configurations, image input dimensions and feature extraction mechanisms described previously in this paper. The influence of additional data augmentation is also shown.

Model	Test AUC	Trainable parameters	Grade
CNN-32×32 (aug)	92.5	2.8 M	6
Caps-32×32-64f-8c.t. (aug)	91.2	1.9 M	8
CNN-32×32	91.7	2.8 M	7
Caps-32×32-64f-8c.t.	90.7	1.9 M	9
Caps-32×32-64f-32c.t.	93.0	3.2 M	2
Caps-32×32-64f-64c.t.	92.6	4.8 M	5
Caps-32×32-512f-64c.t.	53.0	10.5 M	10
Caps-32×32-256f-64c.t.	92.6	7.2 M	4
Caps-64×64-256f-32c.t.	93.2	12.0 M	1
Caps-32×32-64f-32c.t.-16d-32d	92.6	7.1 M	3

Table 7: Performance of different CapsNet and baseline CNN models for the lung nodule candidate patch classification task. The corresponding number of trainable parameters of each network is given in order to show the size of the network, which usually strongly correlates with the computation time required, and may also influence the generalization ability. The grade scale in terms of performance is also presented (lower is better).

Some of the best performing CapsNet models are also presented in the graph of Fig [17], where the corresponding number of filters, A, used in the convolutional layer prior to the capsule layers, the dimensionality of the capsules in *PrimaryCaps* and *ClassificationCaps* layers, C and G, the number of primary capsule types, D, as well as the number of trainable parameters and the test AUC measure on for lung nodule candidate patch classification task are presented.

Once three orthogonal patches are fed to the network for each nodule candidate, and further the prediction for each of these patches is considered as a vote for predicting whether this location is of positive or negative

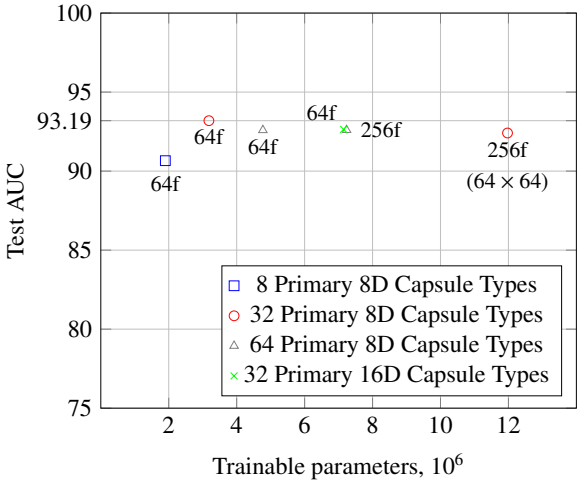


Figure 17: AUC (area under ROC curve) for different configurations of CapsNet models. One may notice that a relatively shallow model with only about 3.2 trainable parameters performs better than the extended configurations with enlarged capsule dimensions of increased capsule types number. The compressed patch size (from the initial 64×64 to 32×32 also does not introduce significant drops in performance. While there is no noticeable improvement in terms of classification AUC when growing different model components, it is sufficient to use a shallow model in that case. The input size I is 32×32 for all models unless otherwise stated.

	Test AUC	
	CNN baseline	**CapsNet**
Separate patches	91.4	93.2
Candidate (patch voting)	93.8	**96.0**

Table 8: Test AUC of the CapsNet and CNN baseline networks for lung nodule classification. The results are presented for classifying separate patches (one of three orthogonal views for each candidate), as well as the result of voting of these three patches orthogonal patches per candidate. The voting among orthogonal views improves the overall accuracy of classification without actually using an ensemble of models.

class, the overall candidate classification accuracy of the networks increases. The result of this voting approach is presented in Tab [8].

5. Discussion

We have confirmed findings in the literature that a capsule network with dynamic routing architecture achieves the state-of-the-art results for a simplistic task like *MNIST* handwritten digit classification. In addition, the capsule network shows much better robustness to affine transformations of data than the best performing traditional convolutional networks. One of the reasons the capsule network works so well with the MNIST data is that it consists of a simple two-dimensional object on

a plain background. This makes it simple for the architecture to learn and route the signals to deeper levels; the network can properly detect, extract and encode the objects present in the image and their instantiation parameters in order to further classify the inputs. This is proved by the quality of the reconstructions obtained from the decoder network.

However, the results are less impressive for more complex data, such as the *smallNORB* dataset. The quality of the encoding and reconstruction was seen to decrease notably. The network is able to extract only the estimated shape of the object for its current appearance to the camera, but no details are preserved. This is due to the increased difficulty of distinguishing the object from the background in this data, where the images are projections of three-dimensional objects; it is much more challenging to estimate their pose and deformations from only the given two-dimensional projections. In order to achieve better estimation of the objects' instantiation parameters, 3D data should alternatively be used.

Despite the above mentioned limitation, our experiments showed that the CapsNet was more stable to the change of viewpoints under which the photos of objects were taken, as shown for the *smallNORB* dataset. However, since this robustness is achieved in a configuration with a rather limited overall classification accuracy, it may be because the network is able to encode only the shape of the object but not its details, it has higher chance of correctly classifying a previously unseen sample than a network which takes into account more minor details.

For the task of lung nodule candidate classification from the false positive reduction task of the *LUNA16* challenge, the CapsNet achieved a 96.0% AUC, compared with the 93.8% of the presented baseline CNN with max-pooling. The mentioned data seems to fit well the architecture of CapsNet while resulting in a sufficiently detailed encoding of the patch features due to being less complex than the samples of *smallNORB* in terms of lighting conditions, reduced background complexity, absence of shadows and greater general consistency. Nevertheless, initially a higher increase in terms of classification accuracy has been expected since unlike for distinguishing simple objects, the presence of certain structures and their orientation relative to the nodule blob was expected to influence the ability to distinguish between the two classes more precisely. More complex statistical patterns were expected to be learned by the CapsNet due to its enhanced preservation of information about the features of the input in deeper layers compared to a CNN with a primitive routing algorithm. However, due to the dominant sphericity of the nodule blobs, the robustness to different variations of its shape appears to be only a minor advantage. Besides, the nodule volume and its size itself remains the most descriptive parameters of the image and the ability to base the network predictions also on the spatial relationships between other objects and the blob introduces only a slightly higher result. After all, taking into account that the described capsule network can be considered as a primary step toward the development of this concept, the achieved performance is satisfying.

In terms of learning ability and the amount of training data required to sufficiently train the network, CapsNet did not show any improvement compared with a conventional CNN. In the study of the network performance versus the training set size, both networks showed the same tendency (i.e. more data resulted in better results). This suggests that we still require enormous datasets to properly train deep neural networks, and there remains a significant difference in the number of training examples required by human learners and machines.

Due to a significantly more complex data routing algorithm, CapsNet has a strong disadvantage when it comes to the computation time required for each training step. Even though both networks require a similar number of iterations to achieve convergence, the capsule network requires approximately eight times more computational power than the baseline CNN network. This is expected, and due to the use of an advanced routing algorithm that requires multiple iterations for every single data pass in order to converge.

The CapsNet and the baseline CNN appeared to have a near equal stability to both CT scanner calibration variance and different CT reconstruction kernel sharpening, while suffering a similar reduction in classification accuracy for the realistic changes of the data.

Modifications of the CapsNet that involved enlarging the networks' size by stacking more capsule layers, increasing the dimensions of capsules in order to allow encoding deformations of higher degrees of freedom, adding convolutional layers for deeper feature extraction, etc. did not seem to affect notably the performance. This leads to the conclusion that the the given configuration is optimal, and that further improvement is limited conceptually rather than computationally. The lack of enforcement in the system that would ensure that the advanced structuring of neural groups correlates to certain physical operations with the features of objects (structured pose and presence prediction) becomes an issue upon the increase of the network size. In the last case the effect from structuring and routing between capsules vanishes and leads to the network acting as a conventional CNN.

One of the largest limitations of the current capsule network architecture is the relatively small input image size that the network can handle. For images of 64×64 pixels the number of parameters in a relatively shallow model could easily reach 15 million, while for bigger and more complex data, such as chest X-rays of size 512×512,

the number of parameters could exceed 100 million. It is worth mentioning that for such complex data, as used in the work of Rajpurkar et al. (2017), where the differences between classes are very minor, the presented capsule network failed to achieve convergence. Therefore, for such tasks as this, where patch based approaches cannot be used due to the way the data is labeled, CapsNet cannot be successfully applied at the moment.

Hinton et al. (2018) have managed to decrease the number of trainable parameters dramatically in the network by introducing sharing of weights among different positions of the same capsule type, as well as a technique referred to as "coordinate addition". The different structure of capsule layers, as well as EM routing algorithm, allowed them to achieve state-of-the-art performances on *smallNORB*. This restructuring limits the operations that a certain region of the network can perform, and avoids vanishing of the conceptual grouping of neurons into capsules as in the CapsNet.

A very recent study by LaLonde R. (2018) expanded the use of capsule networks to the task of object segmentation, where they achieved state-of-the-art results for segmenting lung cavities of the *LUNA16* volumes. They significantly enhanced the concept by modifying the original dynamic routing algorithm to act locally when routing children capsules to parent capsules, and to share transformation matrices across capsules within the same capsule type. These changes dramatically reduce the memory and parameter demand of the original capsule implementation, and allows for operating on large image sizes. To compensate for the loss of global information, they introduce the concept of a "deep convolutional-deconvolutional capsule architecture".

6. Conclusions

Conventional CNNs suffer from a number of drawbacks such as poor generalization, huge training data requirements, low stability to rotational and other geometrical distortions and lack of spatial dependencies and neuron hierarchies. Some of these issues are tackled by the novel capsule network architecture and the dynamic routing algorithm. During this work we have investigated this innovative approach.

The performance of the implemented network has been verified on *MNIST* data, that allowed to prove the implementation correctness and confirm some of the claimed advantages of the CapsNet. In order to study the ability of this architecture to process data of higher complexity, its performance for classification of small images containing 3D objects has been studied. We have shown that CapsNet was more robust to the change of camera viewpoints than a conventional CNN with max-pooling, however the overall accuracy couldn't reach the claimed value.

We proposed a solution for lung nodule candidate classification based on the novel capsule network architecture and the dynamic routing algorithm. It has proven to be more accurate in terms of AUC than the baseline CNN approach, which may be explained by the enhanced detail preservation in deeper layers of the network. The *LUNA16* candidate patches are well processed by this architecture, while the network is able to properly encode and recreate the most important input details and, therefore, produce accurate class predictions. The ability to utilize the knowledge about spatial relationships between anatomical structures present on the patches due to the networks' design, undoubtedly, enhances the ability to distinguish the classes. However, knowledge of additional minor features, that may be further achieved by improved capsule architectures, would probably allow to improve the classification results, while creating a more advanced feature space for which statistical patterns of greater precision could be learned. The way in which the capsule vectors encode information regarding the input features has been illustrated as well as the response of the decoder network to minor changes of this compressed capsule representation. Due to the lack of logical learning constraints for capsules, physical properties that they encode are often redundant. The last should not be considered as a major issue, however improved structuring and neuron group hierarchies might result in better ability to adapt to novel data.

We have also conducted various experiments that allow to better understand the current stage of the concept, its advantages and possible improvements. A strong limitation has been noticed as for the input size and overall data complexity, that the network can properly manage. A data of higher entropy and size introduces a need of certain logical enforcement that would manage different groups of neurons to act as units with defined physical meaning such as feature pose, deformation and presence estimation. Data of growing size steeply increases the number of trainable parameters of the network and in order to be able to train deeper models, has to be reduced.

It was noted that the current capsule network architecture requires significantly bigger computation time for training. We believe this issue might be solved in the future by reduction of trainable parameters in the network, while enhancing its structure as recently presented in the work of LaLonde R. (2018).

Since the concept is being under an early stage of development, some limitations take place such as low scalability, the inability to process data of bigger size and increased complexity. However, unequivocally, some of the current limitations will be solved in the future, creating a powerful alternative to conventional convolutional neural networks.

7. Acknowledgments

This work would not have been possible without valuable supervision, influence and scientific advice of Keith Goatman from Canon Medical Research Europe.

This project was supported and funded by the Education, Audiovisual and Culture Executive Agency (EACEA) as part of Erasmus Mundus master program in Medical Imaging and Applications (MAIA).

I would also like to show my gratitude to the Canon Medical Research Europe company at Edinburgh, United Kingdom for providing excellent placement, top-level computational equipment and access to various datasets for the timespan of the project.

References

Cheng, C.W., Zhao, L., Wolanski, M., Zhao, Q., James, J., Dikeman, K., Mills, M., Li, M., Srivastava, S.P., Lu, X.Q., Das, I.J., 2013. Comparison of tissue characterization curves for different ct scanners: implication in proton therapy treatment planning. Translational Cancer Research 1. URL: http://tcr.amegroups.com/article/view/811.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115 EP –. URL: http://dx.doi.org/10.1038/nature21056.

Gallardo-Estrella, L., Lynch, D.A., Prokop, M., Stinson, D., Zach, J., Judy, P.F., van Ginneken, B., van Rikxoort, E.M., 2016. Normalizing computed tomography data reconstructed with different filter kernels: effect on emphysema quantification. European Radiology 26, 478–486. URL: https://doi.org/10.1007/s00330-015-3824-y, doi:10.1007/s00330-015-3824-y.

Gierada, D.S., Bierhals, A.J., Choong, C.K., Bartel, S.T., Ritter, J.H., Das, N.A., Hong, C., Pilgram, T.K., Bae, K.T., Whiting, B.R., Woods, J.C., Hogg, J.C., Lutey, B.A., Battafarano, R.J., Cooper, J.D., Meyers, B.F., Patterson, G.A., 2010. Effects of ct section thickness and reconstruction kernel on emphysema quantification: Relationship to the magnitude of the ct emphysema index. Acad Radiol 17, 146. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2818169/, doi:10.1016/j.acra.2009.08.007. 19931472[pmid].

Goodfellow I., S.J.S.C., 2014. Explaining and harnessing adversarial examples. ArXiv e-prints arXiv:1412.6572.

Hinton, G.E., Sabour, S., Frosst, N., 2018. Matrix capsules with em routing. International Conference on Learning Representations URL: https://openreview.net/forum?id=HJWLfGWRb.

Kurakin A., Goodfellow I., B.S., 2016. Adversarial examples in the physical world. ArXiv e-prints arXiv:1607.02533.

LaLonde R., B.U., 2018. Capsules for object segmentation. ArXiv e-prints arXiv:1804.04241.

LeCun, Y., Huang, F.J., Bottou, L., 2004. Learning methods for generic object recognition with invariance to pose and lighting, in: Proceedings of CVPR'04, IEEE Press.

Li, K., Yip, R., Avila, R., Henschke, C.I., Yankelevitz, D.F., 2017. Size and growth assessment of pulmonary nodules: Consequences of the rounding. Journal of Thoracic Oncology 12, 657 -- 662. doi:https://doi.org/10.1016/j.jtho.2016.12.010.

Perandini S., Soardi G., M.M.O.E.Z.L.M.S., 2016. Distribution of solid solitary pulmonary nodules within the lungs on computed tomography: A review of 208 consecutive lesions of biopsy-proven nature. Polish Journal of Radiology doi:10.12659/PJR.895417.

Quinn Colin Meisinger, Jeffrey S. Klein, K.J.B.G.G.B.J.L., 2011. Ct features of peripheral pulmonary carcinoid tumors. American Journal of Radiology doi:10.2214/AJR.10.5954.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M.P., Ng, A.Y., 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. ArXiv e-prints arXiv:1711.05225.

Sabour, S., Frosst, N., E Hinton, G., 2017. Dynamic routing between capsules. ArXiv e-prints arXiv:1710.09829.

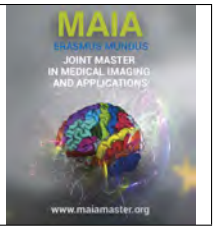
Wan, L., Zeiler, M., Zhang, S., Cun, Y.L., Fergus, R., 2013. Regularization of neural networks using dropconnect, in: Dasgupta, S., McAllester, D. (Eds.), Proceedings of the 30th International Conference on Machine Learning, PMLR, Atlanta, Georgia, USA. pp. 1058--1066. URL: http://proceedings.mlr.press/v28/wan13.html.

Xi E., Bing S., J.Y., 2017. Capsule network performance on complex data. ArXiv e-prints arXiv:1712.03480.

Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O., 2016. Understanding deep learning requires rethinking generalization. ArXiv e-prints arXiv:1611.03530.

Medical Imaging and Applications

Master Thesis, June 2018



Application of machine learning techniques for classification of Healthy controls from patients with Schizophrenia and Bipolar Disorder based on MRI

Benjamin Lalande Chatain, Mariano Cabezas, Arnau Oliver, Xavier Lladó

Dept of Computer Technology and Architecture, University of Girona

Abstract

The diagnosis of psychiatric disorders is currently based purely on the apparition of clinical manifestations. Nevertheless, such a diagnosis reaches its limits during early stage of the development of psychiatric disorders conducting to a follow up of several years to get a first diagnosis in the most extreme cases. The purpose of our study is to develop an automatic aided diagnosis tool to help to classify patients with schizophrenia or bipolar disorders from healthy subjects based on the brain morphology. An analysis of the impact of the development of psychiatric on subcortical structures is conducted. In addition, 5 models are trained both in local classification using support vector machines and global classification using convolutional neural network on a dataset consisting of 125 healthy control subjects, 50 patients with schizophrenia and 49 patients with bipolar disorders. Using convolutional neural network, we achieve an accuracy of 0.89 for the classification of healthy controls and non healthy patients (including both schizophrenia and bipolar disorders) and 0.86 in the subsequent classification of schizophrenia and bipolar disorders patients. In summary, this work demonstrates that analyzing brain morphology we are able to discriminate between normal subjects and schizophrenia and bipolar disorders patients.

Keywords: MAIA Master, Schizophrenia, Bipolar Disorder, Classification, Support vector Machine, Deep Learning, Structural Neuroimaging

1. Introduction

Psychiatric disorders remain a recent field in the scientific community with the term schizophrenia emerging in 1910 by the Swiss psychiatrist Paul Eugen Bleuler, while bipolar disorder, previously known as manic-depressive illness, became only official in the late 90'. Despite their recent recognition, they already represent non-officially 1% and 5% of the global population, respectively, and reach the top 10 of the most disabling diseases by the World Health Organization. Their position may be explained by their high social impact in the daily life of the patient as well as his/her family circle and also their high level of suicide, especially in bipolar disorder patients, where the statistics indicate that 1 patient over 5 will make a suicide attempt during their life.

Currently, the diagnosis of psychiatric disorders such as schizophrenia and bipolar disorders is limited to the apparition of clinical manifestations. Such a diagnosis is efficient at medium or late stage of the disorder when the symptoms appear clearly. Nevertheless, during the early stage of the development of the psychiatric disorder, providing a diagnosis based purely on clinical manifestations reaches its limits due to the low discrimination between the symptoms which can belong to other psychiatric disorders.

Studies regarding the classification of psychiatric disorders based on brain morphology using machine learning techniques emerged with Davatzikos (2005) using Support Vector Machines (SVM) as the main strategy. Koutsouleris (2009) and Ingalhalikar (2010) have continued with this strategy, respectively for the classification of early stage of schizophrenia, and advanced stage of schizophrenia and autism spectrum disorders. An alternative to SVM for classification is the use of Discriminant Function Analysis introduced by Karageorgiou (2011) and Kasparek (2011) for the classification of patients with schizophrenia. To resolve the issue

of detection of psychiatric disorders during early stage, several studies have been published (Strakowski, 2000, 2005), highlighting a correlation between the development of psychiatric disorders and their impact on structural neuroimaging, especially in the subcortical structures, which appear to be enough discriminative to be use as features for classification (Heckers, 2001).

The medical image modality selected to perform the classification is the Magnetic Resonance Imaging. MRI does not involve exposure to radiation and is efficient to show soft tissues structures (i.e. brain tissues). Furthermore, specific type of MRI such as functional MRI can provide information about the blood circulation.

In this master thesis we aim to develop an automatic aided diagnosis tool to perform the classification of psychiatric disorders such as schizophrenia and bipolar disorders based only on neuroimaging information and applicable in early stage of development. To that aim, we have:

- reviewed the impact of psychiatric disorders in the development of the brain morphology,
- reviewed the state of the art regarding the classification of psychiatric disorders using neuroimaging,
- segmented the subcortical structures using multi-atlas based Segmentation and FIRST based Segmentation from the FSL toolbox,
- studied the volumes of different subcortical structures in normal subjects and in patients with psychiatric disorders,
- classified schizophrenia and bipolar disorders patients using SVM based on features extracted from the subcortical structures,
- classified schizophrenia and bipolar disorders patients using convolutional neural network (CNN)
- analyzed the advantages and drawbacks of the above strategies.

The rest of this paper is structured according to the above steps.

2. State of the art

In this section, we will introduce the state of the art from neuroimaging and medical imaging perspectives.

2.1. Review of structural neuroimaging

Besides the recent increase on literature regarding the impact of psychiatric disorders of the brain, the heterogeneity of the methodology and the different clinical populations appear to be a restraint to the comprehension of the role of structural neuroimaging due to the conflicts and contradictions between the different publications.

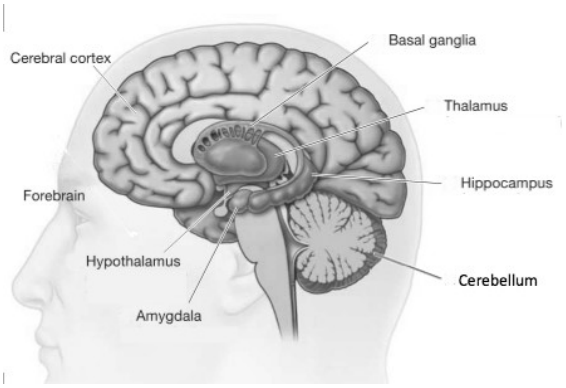


Figure 1: Location of the subcortical structures between the cerebral cortex and the cerebellum

A summary regarding the global and local brain abnormalities with an important impact regarding psychiatric disorders is presented on the table 1. A complete review has been published by Emsell (2009). In next subsections we summarize the main findings of that work.

2.1.1. Global structural abnormalities

Despite the preservation of global volumes in bipolar disorder in contrast to schizophrenia, as documented by Hoge (1999), several publications support the presence of regional deficits in both gray and white matter. Nevertheless, many studies appear to enter in conflict regarding the abnormalities in gray and white matter. Lyoo (2004) demonstrated regional volume reduction of gray matter in multiple prefrontal areas while Lochhead (2004) reporting both regional increases and decreases.

Regarding white matter, the most consistent abnormalities detected are the presence of white matter hyperintensities on T2-weighted and FLAIR MRI on patients with bipolar disorder.

2.1.2. Local structural abnormalities

Although many structures are described having an influence on brain morphology, the literature suggests that subcortical structures have the highest impact due to their cognitive functions. These subcortical structures, consisting of the the Basal Ganglia, the Thalamus, the Amygdala and the Hippocampus, lie directly between the cerebral cortex and the cerebellum (figure 1).

Subcortical structures are important for the classification of psychiatric disorder as their functions have a direct correlation with the behavior of the patient and its symptoms such as the control of the emotions, the sleep or the attention. Namely,

Basal Ganglia: Neurological lesions on Basal Ganglia can implicate on Obsessive Compulsive Disorder

Table 1: Review of structural neuroimaging regarding the presence of local and global abnormalities on patients with psychiatric disorders

Structures	Affectation
Global Volumes	Preservation in Bipolar disorder in contrast to Schizophrenia
White matter	Presence of white matter hyper intensities on T2 weighted Sparse volume reduction in Bipolar disorder
Basal Ganglia	Enlargement of the striatum in Bipolar disorder
Thalamus	Thalamic deficits in Schizophrenia Volume preservation in Bipolar disorder
Amygdala	High heterogeneity Various volumetric changes
Hippocampus	Bilateral volumetric decrease in Schizophrenia Volumetric preservation in Bipolar disorder

and Anxiety disorder. An enlargement of the striatum, the largest component of the subcortical structure, have been put in evidence by Aylward (1994) and DelBello (2004) while Beyer (2004) reports small volumetric decreases for patients with bipolar disorders

Thalamus: Most of the studies report a preservation of the volume in bipolar disorder ((Caetano, 2001; DelBello, 2004)). Inversely, thalamic volume in schizophrenia is repeatedly affected by deficits.

Amygdala: Amygdala dysfunction is subject to conflicts in the scientific community. Indeed there are many publications reporting both volumetric increases and decreases with very high heterogeneity regarding the patient for schizophrenia and bipolar disorder

Hippocampus: The majority of the studies report a bilateral volume deficit ((Videbech, 2004; Wright, 2000)) with a tendency to the left side in schizophrenia while it appears to be preserve in bipolar disorder.

2.2. *Magnetic Resonance Imaging*

Magnetic Resonance Imaging is commonly used for neuroimaging due to its high resolution and its non exposure to radiations. It exists several type of MRI which are presented on figure 2.

The most common MRI are the T1-weighted and the T2-weighted. Their difference are the Repetition Time (TR) and the Time to Echo (TE), the T2-Weighted presenting longer TR and TE, and modifying the contrast and brightness of the tissues.

An another common MRI is the Fluid Attenuated Inversion Recovery (Flair) which has longer TR and TE than the t1-Weighted and T2-Weighted which has the advantage to highlight abnormalities.

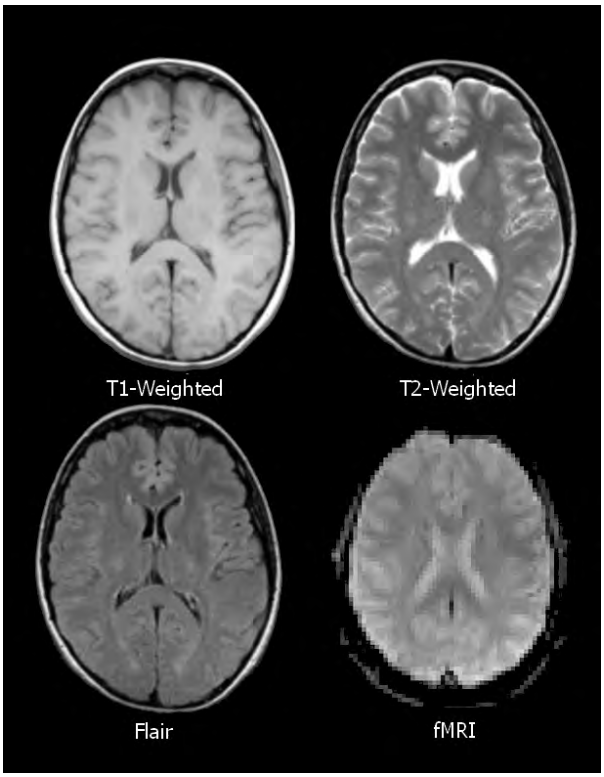


Figure 2: Different Magnetic Resonance Imaging modalities

The last type of MRI presented is the functional MRI which represent the blood oxygen level dependent (BOLD) signal, traducing the cerebral activity.

2.3. *Classification of patients with schizophrenia or bipolar disorders*

As mentioned, despite the growing interest regarding the classification of psychiatric disorders in medical imaging, the recentness of its recognition by the scientific community leads to a limited state of the art. To the best of our knowledge, only a couple of studies compose the state of the art for the classification of schizophrenia and bipolar disorder, the most recent being Nieuwenhuis (2012) and Schnak (2013), both works using SVM on structural 3T MR images.

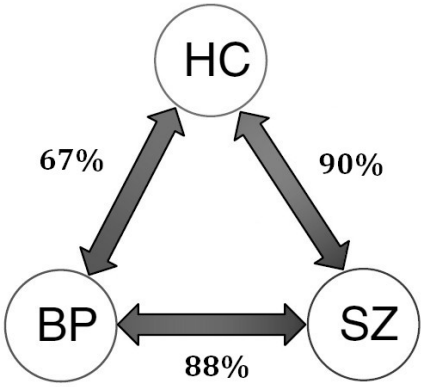


Figure 3: Score of the state of the art using the Support vector Machine models based on gray matter densities

Both publications are established on the assumption of the presence of brain abnormalities on patient with psychiatric disorder ((Arnone, 2009; Ellison-Wright, 2010; Hulshoff, 2012; McDonald, 2005)) represented by local presence, or concentration, of gray matter defined as gray matter densities (GMDs) in specific neuroimaging structures. After transformation into a standardized coordinates system of the GMDs images to GMDs maps, 3 independent models are built using the machine learning technique SVM. The 3 independent models, entitled later as the three two-class classification due to their triangular aspect (figure 3), classify respectively healthy controls from schizophrenia, healthy controls from bipolar disorder, and schizophrenia from bipolar disorder.

On the latest publication (Schnak (2013)), the schizophrenia patients were classified from the healthy control subjects with an accuracy of 90%. The schizophrenia patients and patients with bipolar disorder were distinguished with an accuracy of 88%. Besides, the classification of patients with bipolar disorder from healthy control patients, which appear as the most difficult, provided the lowest accurate model with an accuracy of 67%. Nevertheless the general classification of non healthy control patients, including both schizophrenia and bipolar disorder, from healthy control subjects, and the 3 classes classification healthy control from schizophrenia from bipolar disorder have not been experimented.

3. Material and methods

In this section we will present the different selected databases and the procedures applied in order to quantify the brain abnormalities on patients affected by mental disorders and to provide a classification.

During the master thesis, the experiments have been conducted using two independent public databases, the Hammers Adult Database and the UCLA Database, which are explained in what follows.

3.1. Hammers Adult Database

The first database is the Hammers Adult database include 30 individual Atlas consisting of a MRI and its associated labeled segmentation. It also provide the adult probabilities atlas for each of the 95 brain structures and the associated regional probabilistic maps

The Hammers Adult database is available in open source[1] and shared by the Faculty of Medecine Imperial College London[2], the University College of London[3] and the Hospital Neurologique Pierre Wertheimer of Lyon[4].

The database is a contribution of Hammer (2003) for the segmentation of the regions of interest 01 to 49, Gousias (2008) for the regions of interest 50 to 83 and Faillenot (2017) for the regions of interest 86 to 95.

3.2. UCLA Database

The second database used in this work, also available as opensource, has been published by the UCLA Consortium for Neuropsychiatric Phenomics LA5c Study[5] (Poldrack (2016) and Gorgolewski (2017)). The database consists of 150 men and 115 women divided into 4 different subset: 125 Healthy control (HC), 50 patients with Schizophrenia (Sz), 49 patients with Bipolar disorder (Bd) and 41 patients with Attention deficit hyperactivity disorder (ADHD). Each patient's folder contains both anatomical and functional MRIs.

3.2.1. Anatomical MRI

The anatomical MRIs correspond to a series of high resolution volumes representing the preprocessed T1-Weighted images, the brain mask and the independent segmentation of the 3 main brain tissues: white matter, gray matter and cerebro-spinal fluid (figure 5). The preprocessed T1-Weighted includes the following preprocessing steps: correction of the bias field, skull-stripping and registration to the Montreal Neurological Institute (MNI) space.

In our work, both segmentation and classification of the subcortical structures will be performed using the anatomical MRI of the UCLA database.

3.2.2. Functional MRI

The functional MRI measures the brain activity by detecting changes of the blood oxygen level dependent (BOLD) signal over the time. The regions with high cerebral activity appear brighter than the rest.

[1] brain-development.org

[2] Department of Clinical Neuroscience and MRC Clinical Sciences Center, Division of Neuroscience and Mental Health

[3] Department of Clinical and Experimental Epilepsy, Institute of Neurology

[4] Functional Neuroimaging, Foundation Neurodis

[5] https://openneuro.org/datasets/ds000030/versions/00002

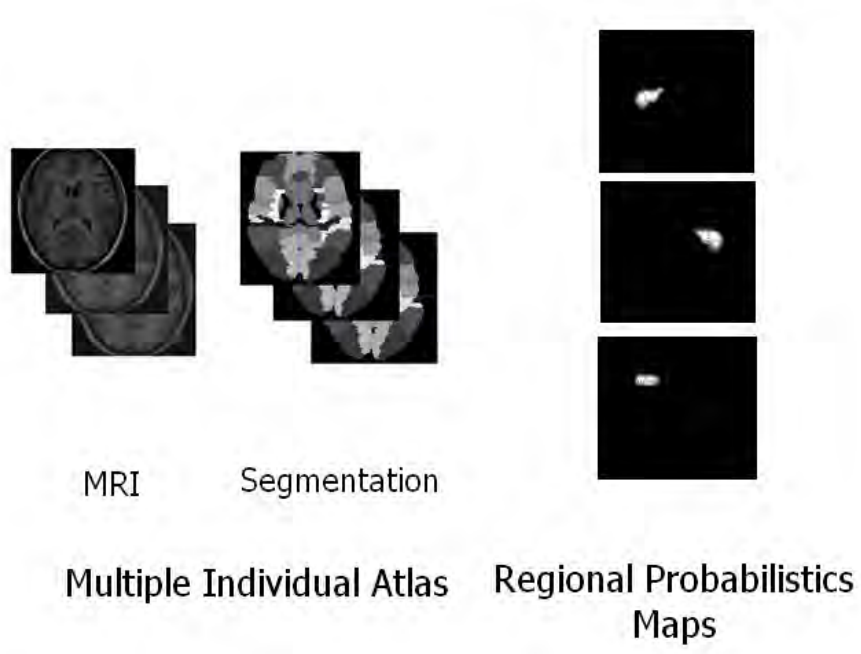


Figure 4: Description of the components of the Hammer Adult database

The series consists of 6 fMRIs T2-weighted, each acquired after a questionnaire and a different neurocognitive task. The following descriptions are the original description of the neurocognitive tasks (extracted from Gorgolewski (2017)):

- Rest: A resting state session eyes open
- Bart: Balloon analog risk task. Participants were allowed to pump a series of virtual balloons. Experimental balloons (green) resulted either in an explosion or in a successful pump (no explosions and 5 points). Control balloons (white) did neither result in points nor exploded. Participants could choose to not to pump but to cash out and start with a new balloon.
- Scap: Spatial working memory task. Subjects were shown an array of 1, 3, 5 or 7 circles pseudorandomly positioned around a central fixation cross. After a delay, subjects were shown a green circle and were asked to indicate whether the circle was in the same position as one of the target circled. In addition to the memory load, the delay period was manipulated with delays of 1.5, 3 or 4.5 seconds. Half of the trials were true positive and half were true negative.
- Stop-signal: Stop signal task. Participants were instructed to respond quickly when a 'go' stimulus was presented on the computer screen, except on the subset of trials where the 'go' stimulus was paired with a 'stop' signal. The 'go' stimulus was a pointing arrow, a stop-signal was a 500 Hz tone presented through headphones.
- Task-switch: Task-switching task. Stimuli were shown varying in color(red or green) and in shape(triangle or shape). Participants were asked to respond to the stimulus based on the task cue (shape 'S' or color 'C'). The task switched on 33% of the trials
- Bht: Breath holding task. Participants were asked to alternate between holding their breath and breathing regularly while resting.

For each description of the neurocognitive tasks above, we highlight specific aptitudes. Indeed, the Balloon analog risk task evaluates the risk perception of the patient while the Breath holding task required patience and self control. The task Switching-tasks and the Stop signal task reward respectively a good visual and sound reaction. The spatial working memory task focuses on visual memory. Finally the Rest task, which has no specific requirements is defined as the reference task.

3.3. Healthy and patients classification using SVM

In this section we will detail the different methods for segmentation of the sub cortical structures and the classification of Healthy control subjects from patients with psychiatric disorders using support vector machine and based of the brain volume abnormalities.

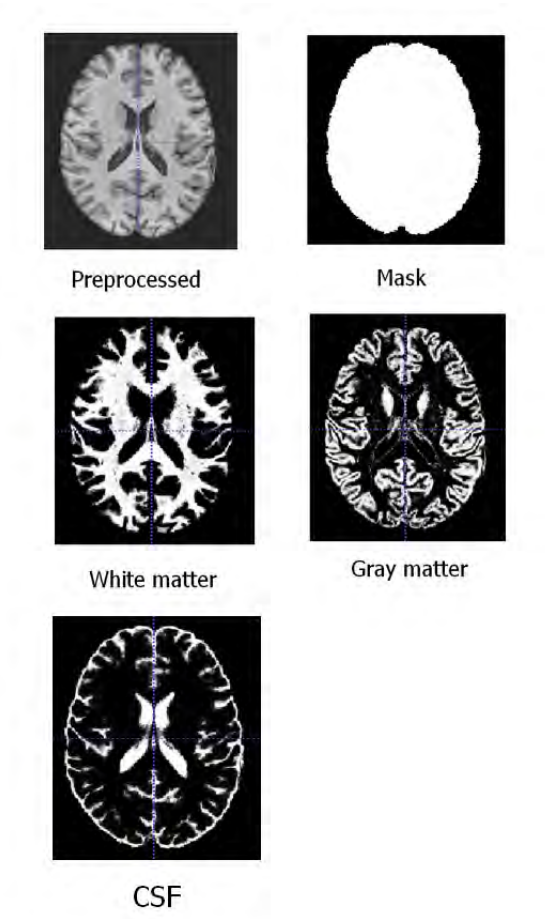


Figure 5: Description of the components of the UCLA database: the preprocessed T1-Weighted, the brain mask, and the 3 independent tissues segmentation: White matter, Gray matter and CSF

3.3.1. Multi-atlas based Brain Structures Segmentation

Multi atlas based segmentation is a common approach in medical image segmentation ((Ltjnen, 2010; Wang, 2012)). It consists on registering each MRI of the Individual Atlas to the target image. Then the labeling decision fusion is done using the Mutual Information to quantify the similarity between the individual atlas's MRI and the target image (figure 7). After defining the best match between the multiple individual atlases and the target image, the same transformation matrices are applied to the associated labeled segmentation. The final output segmentation is a 3D volume labeled with 95 brain regions.

The multiple individual atlases of the Hammers Adult database are registered to each of the preprocessed T1-Weighted images of the UCLA database.

The registration is achieved using the open source libraries NiftyReg, developed at University College London, which perform successively affine and non-rigid registration.

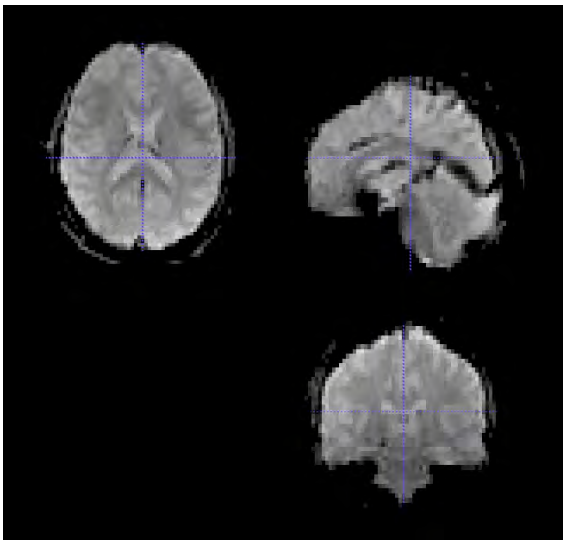


Figure 6: Preview of functional MRI modality.The bright region represent the regions where the cerebral activity is at the highest

3.3.2. FIRST segmentation

The second segmentation of the brain structures is performed using the FIRST model-based segmentation and registration tool from the FSL toolbox of analysis tools developed by the University of Oxford, especially introduced by Patenaudeh (2011) and detailed in his thesis (Patenaudeh, 2007), containing more technical details.

The FIRST algorithm provide an automatic segmentation for 15 brain regions, including sub cortical regions, based on atlas segmentation approach, searching through linear combinations of shape modes of variation for the most probable shape instance,based on multivariate Gaussian assumptions, given the observed intensities in a T1-weighted image.

3.3.3. Classification using Support Vector Machine

Once we have the structures segmented, a SVM discriminative classifier is considered to build our supervised learning model. The SVM is defined as a hyperplane or set of hyperplanes in a high dimensional space. The classification is based on the distance to the nearest training-data point of any class.

The classification is performed by extracting of the local features of each subcortical structures previously segmented. The main feature for each structure is their volume in voxels normalized by the whole brain volume.

Nevertheless, limiting the features to the unique volume of the subcortical structures is not discriminative enough. That is the reason why additional features have been given to the SVM classifier such as the mean intensity and the standard deviation of the subcortical structure.

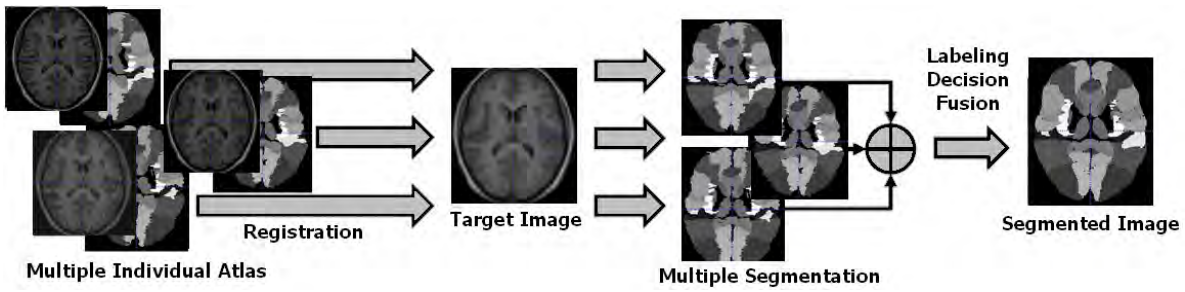


Figure 7: Step by step method of multi-atlas based segmentation

3.4. Healthy and patients classification using Convolutional Neural Networks

In contrast with SVM that are feeded with handcrafted features (i.e. the volumes of the structures), CNN automatically extracts the best features during training. Hence, the input to the CNN is already the images. In what follows we explain this approach in more detail.

3.4.1. Convolutional Neural Network

During years, conventional machine learning techniques have been the standard for classification until the emergence of deep learning techniques based on CNN. The recent investigations on that field, accompanied with the progress in GPU development, makes CNN the current state of the art technique in machine learning.

A CNN is defined as a succession of layers, each of them having a different function:

- Convolutional layer: The convolutional layer is the main layer of a CNN architecture. They are stack one after the other, and can be seen as a pyramid. Each convolutional layer is compose of a set of filters where each kernel is slided over the input image to extract features. The first convolutional layers return low level features while the last ones return more complex features.

- Pooling layer: The most common Pooling Layer is the Max Pooling Layer. It allows to reduce the number of parameter on the next layers while control the over fitting. The output of the Max pooling layer correspond is the max of each region represented by the filter.

- Fully connected Layer: The fully connected layer, or dense layer, is unique by its characteristic to has a full connection to all the activations of the previous layer. The ouput of the fully connected is the predicted class labels.

- Dropout layer: Dropout layer is a regularization layer for reducing overfitting in neural networks by preventing complex co-adaptations on training data. During the dropout layer, a partition of the neurons are deactivated to force the layer to learn the same concept from different neurons

The full architecture of the CNN used is depicted on figure 9. The classification of psychiatric disorders has been implemented using the Keras, a high-level neural networks API[6]

3.4.2. Transfer Learning

Transfer Learning consists of employing a model trained for a specific classification task along with its learned weights and use them for another task. Transfer learning is an optimization step that allows rapid progress. Its general use is in the presence of a small database because deep net do not train well with a small number of samples. The trained model selected is the VGG-16 developed by the University of Oxford for the International Classification Challenge ILSVRC-2014 (Simonyan (2014)).

3.4.3. Fine Tuning

Fine tuning the CNN is a common strategy to continue to fine-tune the weights of the pretrained network by continuing the backpropagation.
It is explain by the presence of more generic features in the first layers (i.e. the pretrained network) and more complex and discriminative features in the last layers on which we want to focus for the classification.

During the process, we applied different techniques of fine tuning to the convolutional neural network to optimize it:

- Freezing the first few layers issues from the pretrained model to preserve their weights because the first few convolutional layers capture low level features which can be re-use in our task.

- Adding of a classifier on the top of the convolutional base by adding a fully connected layer followed by a softmax layer with the number of classes of the task as parameter.

[6]The documentation relative to Keras is available at https://keras.io/

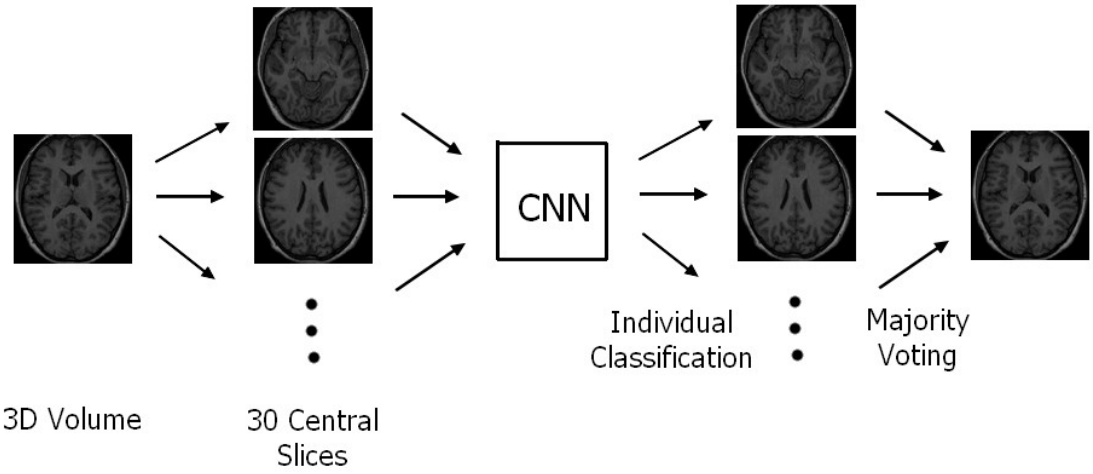


Figure 8: Procedure of data processing for convolutional neural network

- Choosing of a smaller learning rate, usually 10 times smaller to preserve them to distortion assuming that the weights of the pre-trained model are quite good.

3.4.4. Procedure

To process the data for the CNN, it is possible to give as input the 3D volume directly or, in contrast, feed it using extracted 2D images from a 3D volume. This latter option has been selected and the full procedure is presented on figure 8. The same procedure has been applied to the training, validation and testing datasets.

From each volume 30 central slices of the Axial view are extracted and given as input to train the CNN. The output is the independent prediction for each central slice of each 3D volume. The final prediction is subsequently obtained by applying majority voting to the predictions of the slices. The choice of the parameters such as number of central slices and axis will be discussed in the section 5.

3.5. Classification of functional MRI using Convolution Neural Network

As mentioned in the section 3.2.2, the UCLA provides in addition of the T1-Weighted a series of functional MRI acquired after the execution of a neurocognitive tasks. In this master thesis we also want to test if using this information can discriminate between the healthy subjects and the patients.

Notice that using functional MRI is slightly different that normal T1-Weighted, since it is defined as a 4D volume, the last dimension being the acquisition time during which the cerebral activity is measured as the blood oxygen level dependent (BOLD). Several techniques have been published to use functional MRI based on signal analysis (Bandettini, 1993) or independent components analysis (Calhoun, 2001).

To highlight the brain areas where the cerebral activity is the highest and correspond to the target brain area of the cognitive test, appearing as brighter on the functional MR images, we choose to compute the average of the 4D volume over the time into a 3D volume. Then the procedure to process the data remains the same than explained previously.

4. Results

First we present the results of the segmentation structures and the differences we found between healthy subjects and patients. Afterwards, the classification results obtained when analyzing the anatomical images using either the SVM or the CNN approaches. Finally, we show the results obtained when feeding the CNN with the functional MR images.

4.1. Segmentation of subcortical structures

4.1.1. Multi-atlas based segmentation

As mentioned in the section 3.3, the first method of segmentation has been the multi-atlas based segmentation presented on figure 10 which provides 95 brain structures.

We observed that the segmentation is not perfect and presents some under segmentation, visible on the sagittal view. Nevertheless, the subcortical structures, which are the final aim of the segmentation, seem to be well segmented.

Figure 9: CNN Architecture

4.1.2. FIRST Based segmentation

The second method of segmentation is FIRST from the toolbox FSL and provide 15 brain structures (figure 11), including subcortical structures.

The contours of the subcortical structures seem to be better segmented than our multi-atlas based segmentation.

Two different segmentation are generated as output for classification using FIRST: a segmentation of the 15 structures, and a segmentation of the 6 subcortical structures composed of the right and left Amygdala, Hippocampus and Thalamus.

4.2. Volumetric differences of segmentation

After segmentation of the brain into structures, the following subcortical structures: Amygdala, Hippocampus and Thalamus, have been extracted. The average volume for each patient's classes for each sub cortical structures have been normalized to the entire brain volume and are shown on table 2.

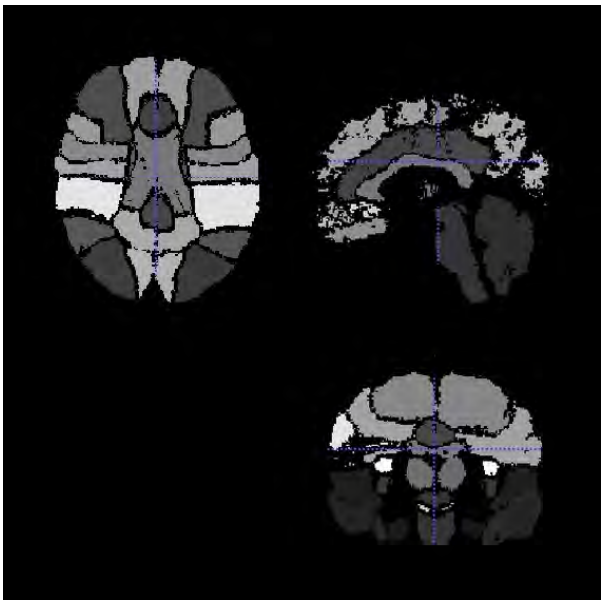


Figure 10: Subcortical structures segmentation using multi-atlas

The two segmentation methods present different results on the same database. The difference is clearly observable with the Hippocampus (figure 12) which is under segmented when performing our Atlas based segmentation.

Nevertheless, the table confirms the assumptions made on the section 2.1.2 regarding the volumetric changes on the subcortical structures in the case of psychiatric disorders, independently of the method of segmentation. Indeed, we observe a volume preservation for patients with bipolar disorders on the Hippocampus and Thalamus regarding the healthy subjects while patients with schizophrenia shows slight deficits. For the Amygdala, as mentioned in the table 1 of the section 2.1.2, we note various volumetric changes depending of the side, right or left, and the method of segmentation.

4.3. Local Classification using Support Vector Machine

In this work, we have tested different classification settings:

- The three 2-class classification presented in the state of the art: healthy control from schizophrenia (Hc - Sz), healthy control from bipolar disorder (Hc -Bd), schizophrenia from bipolar disorder (Sz - Bd)
- The 2-class classification: healthy control from non healthy control, including both schizophrenia and bipolar disorder (Hc - nHc)

Table 2: Volumetric comparison of the subcortical structures depending of the diagnosis

Segmentation	Diagnosis	Amyg. R	Amyg. L	Hipp. R	Hipp. L	Thal. R	Thal. L
Atlas	Healthy	0.1242	0.1214	0.1684	0.1410	0.5682	0.5237
	Schizophrenia	0.1233	0.1212	0.1638	0.1402	0.5283	0.5208
	Bipolar	0.1230	0.1201	0.1661	0.1408	0.5573	0.5240
First	Healthy	0.1106	0.1066	0.2927	0.2789	0.5905	0.6084
	Schizophrenia	0.1078	0.1053	0.2900	0.2720	0.5762	0.6056
	Bipolar	0.1086	0.1057	0.2918	0.2766	0.5893	0.6071

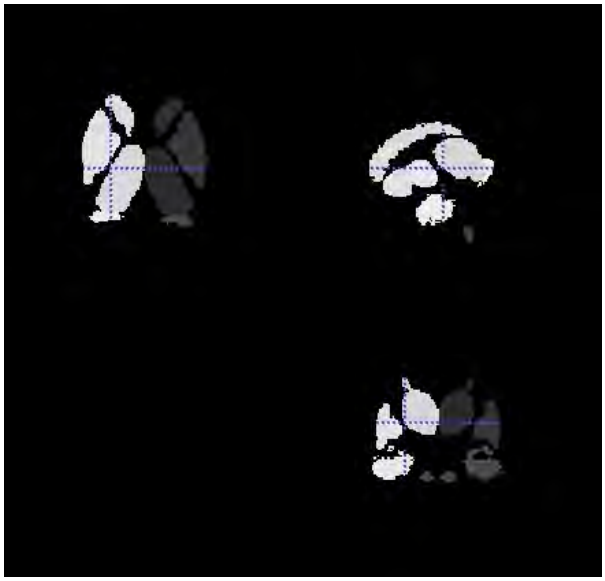


Figure 11: Subcortical structures segmentation using FIRST

Figure 12: Comparison of segmentation methods for subcortical structures: on the left our Atlas based segmentation, on the right the FIRST algorithm

- The 3-class classification: healthy control from schizophrenia from bipolar disorder (Hc - Sz - Bd).

Table 3 shows the results of the above classification settings using SVM. To recall, the features given as input to the SVM are the normalized volume, the mean intensity and the standard deviation of each of the subcortical features computed either using the multi-atlas segmentation or the FIRST algorithm. The classification give the highest results when using the FIRST-based segmentation of the 15 subcortical structures, especially on the classification of healthy control subjects from schizophrenia (+0.05) and the schizophrenia from bipolar disorder (+0.06).

Two observations stand out from the crowd. The first is that the FIRST segmentation provides more accurate segmentation in general of the subcortical structures than our multi-atlas based segmentation. The second is that limiting the segmentation only to the subcortical structures reduces the accuracy. This is explained by the lack of information of the adjacent structures which are directly impacted by the volumetric changes of the subcortical structures.

4.4. Global Classification using Convolutional Neural Network

Table 4 compare the results obtained with global classification using CNN to the local classification using SVM and the state of the art. Nevertheless, it is important to note that the results from the State of the art have been obtained on a different database and their comparison is purely indicative.

Regarding the 2-class classification, CNN present a major increase of the accuracy from the SVM. The classification of the healthy controls from schizophrenia patients shows an increase of +0.24, schizophrenia from bipolar disorders +0.19, and healthy controls from non healthy controls +0.20. The lowest increase of accuracy is for the classification of healthy control from bipolar disorder(+0.13) which appear to be in adequacy with the results of the state of the art. Indeed from the three 2-class classification, the classification of healthy control from bipolar disorder is shown as the hardest classification.

About the 3-class classification, healthy control from schizophrenia from bipolar disorders, the support vector machine present slight better accuracy than the convolutional neural network accuracy (+0.07).

4.5. Classification of functional MRI using Convolution Neural Network

Finally, table 5 presents the results of the classification of the functional MRI. To recall, a functional MRI tracks the Blood Oxygen Level Dependent (BOLD) signal which represents the cerebral activity of the patient. Each functional MRI has been acquired after a neurocognitive task, described on the section 3.2.2 and a questionnaire.

Table 3: Local classification using Support Vector Machine in function of the method of segmentation

	Hc - Sz	Hc - Bd	Sz - Bd	Hc - nHc	Hc - Sz - Bd
Atlas 6 Structures	0.52	0.51	0.52	0.50	0.36
First 6 Structures	0.60	0.63	0.61	0.66	0.52
First 15 Structures	0.65	0.64	0.67	0.70	0.52

Table 4: Comparison between the State of the art, the Local classification using Support Vector Machine and the Global classification using Convolutional Neural Network

	Hc - Sz	Hc - Bd	Sz - Bd	Hc - nHc	Hc - Sz - Bd
State of the art	0.90	0.67	0.88		
SVM	0.65	0.64	0.67	0.70	0.52
CNN	0.91	0.77	0.86	0.89	0.45

The original functional MRI being 4 dimensional, the 4D volume has been converted to a 3D volume by doing the average over the time to be given as input on the CNN. The classification has been performed as a 2-class problem between healthy control subjects and non-healthy control subjects, hence grouping both schizophrenia and bipolar disorder patients together.

The Task Rest correspond to the functional MRI of reference which explain the accuracy close to 50%. Regarding the others neurocognitive task, we note that only the task bart (Balloon analog risk task) and bht (Breath holding task) present the highest accuracy. These cognitive test have in common respectively the perception of risk, the reaction to the stress and require patience during the activity which are known to be secondary symptoms of psychiatric disorders, both schizophrenia and bipolar disorders. The rest of the neurocognitive tasks, scap (Spatial memory task), stopsignal (Stop signal task), switch (Switching task) have an accuracy close to 50%. It is explained by the type or neuro-stimulus generated by the cognitive test, respectively spatial coordination and reaction to a visual or sound signal, which do not affect any of the psychiatric disorders studied in this work.

5. Discussion

5.1. Parameters of the input data for Convolutional Neural Network

As mentioned previously, the CNN approach has two distinct parameter defining the format of the input image: the number of central slices and the MRI axis.

For the number of central slices, experiments have been conducted over the range [10 60] without significant changes regarding the accuracy of the predictions, the final parameter has been set to 30, which represent the average and include all 3 subcortical structures.

Regarding the axis, no significant changes have been noticed over the axial, sagittal and coronal axis. By default, the input image has been set to the axis Axial because it is the common acquisition axis for MRI

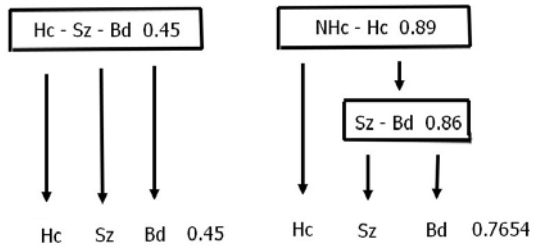


Figure 13: Comparison between 3 classes classification and two successive 2 classes classification

5.2. Limitation of the 3 class classification

As observed on the previous section, there is a clear limitation of the 3 class classification problem (healthy control from schizophrenia from bipolar disorders) with an accuracy of 0.45.

Indeed, we note that the successive application of CNN for 2 class classification problem provide more accurate results (figure 13. By performing first the classification of healthy control from non healthy control (0.89) followed by the classification schizophrenia from bipolar disorders (0.86), we reach similar predictions than the 3 classes classifications with a better accuracy, in theory equal to 0.7654.

5.3. Pros and cons of the integration of functional MRI for classification

The integration of functional MRI for classification of non healthy control patients shows that depending on the neurocognitive task, a classification is possible with an accuracy of 0.685 for the task Bart and 0.64 for the task Bht. Nevertheless, such an accuracy is possible only if the reaction to the neuro-stimulus generated by the neurocognitive task is shared by both of the psychiatric disorders.

An another constraint of the use of functional MRI is the need to register the fMRI to the anatomical MRI for

Table 5: Classification of fMRI using Convolution Neural Network

Cognitive task	Task bart	Task bht	Task rest	Task scap	Task stopsignal	Task switch
Accuracy CNN	0.685	0.64	0.52	0.51	0.545	0.56

information fusion which is difficult regrading the 4th dimensionality of the fMRI.

6. Conclusions

During the master thesis, the role of the subcortical structures has been highlighted during the development of psychiatric disorders such as schizophrenia or bipolar disorders.

A segmentation of the subcortical structures has been performed using our own multi-atlas based segmentation approach and FIRST algorithm from the FSL toolbox. A quantitative study of the subcortical volume has shown brain abnormalities located on these subcortical regions. A local classification using SVM has been conducted based the subcortical volume and has demonstrated that the FIRST segmentation presents a better segmentation than our Multi atlas approach.

Then we performed a global classification using CNN based on the preprocessed T1-Weighted MRI issue from the UCLA database. Performing successive 2 class classification, healthy controls subjects from non healthy control followed by schizophrenia from bipolar disorders, appeared to provide more accurate predictions than the normal 3 class classification, healthy control from schizophrenia from bipolar disorders.

Finally a global classification using CNN based on functional MRI has been achieved. Nevertheless, a classification based on functional MRI has shown that the associated cognitive task is primordial and highly depends on the neuro-stimulus generated.

To conclude, a computer aided diagnosis tool has been develop based on the presence of abnormalities on the sub cortical structures using deep learning techniques, with an accuracy reaching 76.5 % for the classification of healthy control from patients with schizophrenia from patients with bipolar disorders.

7. Acknowledgments

First, I would like to thank my supervisors Arnau Oliver, Xavier Lladó and Mariano Cabezas for giving me the opportunity to work on a subject that concerns me directly, and for their constant follow up during the master thesis. Without it, completing the thesis would not have been possible.

I would like also to thanks the staff of the VICOROB laboratory, Sergi, Jose, Kaisar, for their guidance during the weekly meetings, as well as my friends and coworkers, Roberto, Luca, Sharon, Albert from the master Maia and Vibot.

I would like to thanks all my professors of the Universities of Burgundy, France, of Cassino, Italy, and Girona, Cataluna, Spain for providing a master of excellence from both an academic and a human point of view. Despite the sleepless nights, this master will remain one of my best memories.

I would like to thank and my friends in France, RapowneD, Porcoros, the Duchesse, for supporting me even when I didn't deserve it.

A last and special thank to m family for believing in me all these years, especially my father which is a source of inspiration and the origin of my career in the medical imaging field.

References

Arnone, D., 2009. Magnetic resonance imaging studies in bipolar disorder and schizophrenia: meta-analysis. Br J Psychiatry 195, 194–201.

Aylward, E.H., 1994. Basal ganglia volumes and white matter hyperintensities in patients with bipolar disorder. American Journal of Psychiatry 151, 687–693.

Bandettini, P.A., 1993. Processing strategies for timecourse data sets in functional mri of the human brain. Magnetic Resonance in Medicine 30, 161–173.

Beyer, J.L., 2004. Hippocampal volume measurements in older adults with bipolar disorder. American Journal of Geriatric Psychiatry 12, 613–620.

Caetano, S.C., 2001. Mri study of thalamic volumes in bipolar and unipolar patient and healthy individuals. Psychiatry Research 108, 161–168.

Calhoun, V.D., 2001. A method for making group inferences from functional mri data using independent component analysis. Human Brain Mapping 14, 140–151.

Davatzikos, C., 2005. Whole brain morphometric study of schizophrenia revealing a spatially complex set of focal abnormalities. Arch Gen Psychiatry 62, 1218–1227.

DelBello, M.P., 2004. Magnetic resonance imaging analysis of amygdala and others subcortical brain regions in adolescents with bipolar disorder. Bipolar Disorder 6, 43–52.

Ellison-Wright, I., 2010. Anatomy of bipolar disorder and schizophrenia: a meta-analysis. Schizophr. Res 117, 1–12.

Emsell, L., 2009. The structural neuroimaging of bipolar disorder. International Review of Psychiatry 21, 297–313.

Faillenot, I., 2017. Macroanatomy and 3d probabilistic atlas of the human insula. NeuroImage 150, 88–98.

Gorgolewski, K.J., 2017. Preprocessed consortium for neuropsychiatric phenomics dataset .

Gousias, I., 2008. Automatic segmentation of brain mris of 2-year-olds into 83 regions of interest. NeuroImage 40, 672–684.

Hammer, A., 2003. Threedimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. Human Brain Mapping , 224–247.

Heckers, S., 2001. Neuroimaging studies of the hippocampus in schizophrenia. Hippocampus, special issue schizophrenia 11, 520–528.

Hoge, E.A., 1999. Meta analysis of brain size in bipolar disorder. Schizophrenia Research 37, 177–181.

Hulshoff, H.E., 2012. Overlapping and segregating structural brain abnormalities in twins and with schizophrenia or bipolar disorder. Arch. Gen. Psychiatry 69, 349–359.

Ingalhalikar, M., 2010. Dti diagnostic prediction of a disease via pattern classification. Med Image Comput Comput Assist Interv 13, 558–565.

Karageorgiou, E., 2011. Neuropsychological testing and structural magnetic resonance imaging as diagnostic biomarkers early in the course of schizophrenia and related psychoses. Neuroinformatics 9, 321–333.

Kasparek, T., 2011. Maximum uncertainty linear discrimination analysis of first episode schizophrenia subjects. Psychiatry Res. 191, 174–181.

Koutsouleris, N., 2009. Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. Arch Gen Psychiatry 66, 700–712.

Lochhead, R.A., 2004. Regional brain gray matter volume differences with bipolar disorder as assessed by optimized voxels based morphometry. Biological Psychiatry 55, 1154–1162.

Lyoo, I., 2004. Frontal lobe gray matter densities decreases in bipolar i disorder. Biological Psychiatry 55, 648–651.

Ltjnen, M., 2010. Fast and robust multi-atlas segmentation of brain magnetic resonance images. NeuroImage 49, 2352–2365.

McDonald, C., 2005. Regional volume deviations of brain structure in schizophrenia and psychotic bipolar disorder: computational photometry study. Br J Psychiatry 186, 369–377.

Nieuwenhuis, M., 2012. Classification of schizophrenia patients and healthy controls from structural mri scans in two large independent samples. NeuroImage 61, 606–612.

Patenaudeh, B., 2007. Bayesian statistical models of shape and appearance for subcortical brain segmentation .

Patenaudeh, B., 2011. A bayesian model of shape and appearance for subcortical brain segmentation. NeuroImage 56, 907–922.

Poldrack, R., 2016. A phenome-wide examination of neural and cognitive function .

Schnak, H.G., 2013. Can structural mri aid in clinical classification ? a machine learning study in two independent samples of patients with schizophrenia, bipolar disorder and healthy subjects. NeuroImage 84, 299–306.

Simonyan, K., 2014. Very deep convolutional networks for large-scale image recognition .

Strakowski, S.M., 2000. Neuroimaging in bipolar disorder. Bipolar Disorder, an international journal of psychiatry and neurosciences 2, 148–164.

Strakowski, S.M., 2005. The functional neuroanatomy of bipolar disorder: a review of neuroimaging findings. Molecular Psychiatry 10, 105–116.

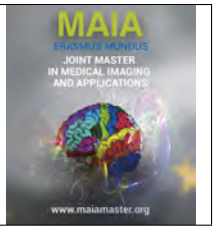
Videbech, P., 2004. Hippocampal volume and depression: A meta-analysis of mri studies. American Journal of Psychiatry 161, 1957–1966.

Wang, H., 2012. Multi-atlas segmentation with joint label fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 611–623.

Wright, I.C., 2000. Meta-analysis of regional brain volumes in schizophrenia. American Journal of Psychiatry 157, 16–25.

Medical Imaging and Applications

Master Thesis, June 2018



False Positive reduction for lesion detection in breast mammography based on two-views lesion correspondence strategy

Maria del Carmen Moreno Genis

Image Analysis Laboratory
VICOROB research institute
University of Girona
Girona, Spain

Abstract

Mammography are used as an efficient tool for breast cancer diagnosis. In the recent years, Computer Aided Diagnosis (CAD) can be very useful for detection of breast cancer. Unfortunately, the presence of high False Positive (FP) detection is an actual issue the majority of CAD systems face. The purpose of this project is the reduction of FP rate of an existed framework for lesion detection. The majority of approaches for FP reduction, and even CAD systems, analyze each breast projection separately. In contrast, our approach is based on a two-views lesion correspondence using a 3D breast deformation and epipolar curves projections. The algorithm proposed faced different dificulties, however the idea of the project have been proven in certain cases. The main drawbacks and the possible factors that cause the algorithm fails are explored as well. As a side of the project, a pectoral muscle segmentation based on deep learning was implemented and different evaluation were made to prove how robust the algorithm can be. Two datasets were using during this project, the INbreast and the Optimam dataset. In total, 234 images were used for the pectoral muscle segmentation section; having as result a Dice Similarity Coeficient of 0.94 and 0.81 for INbreast and Optimam images respectely. In the case of the two-view lesion correspondence, 26 patients containing 52 lesions were examinated.

Keywords: breast cancer, U-net, false positive reduction, digital mammography, curve epipolar line, CAD, deep learning, segmentation.

1. Introduction

Breast cancer is the most common cancer diagnosed among women worldwide (Siegel et al., 2014). According to the American Cancer Society (ACS), breast cancer constitutes 25 percent of all new cancer diagnoses in women. In 2012, almost 1.7 million new cases were diagnosed worldwide (Society, 2013). The mortality of breast cancer can be reduced largely by identifying the cancer at the initial stage. The World Health Organization (WHO) estimates that, although the survival rates for breast cancer vary worldwide, in general, rates have improved in the last years. This was observed by Ferlay et al. (2015), where the observed survival rates were much higher for early stage detected cancers (80-90%) than advanced stage cancers (24%).

X-ray mammography is the gold standard to detect breast cancer in its early stages since decades (Misra et al., 2010). In clinical practice, the conventional mammography exam typically consists on four images: a mediolateral oblique (MLO) and a craniocaudal (CC) views for each breast. The MLO view is a projection taken at an angle of approximately 45 degrees, where, in most of the cases the pectoral muscle is captured. On the other hand, the CC projection corresponds to a top-down view of the breast (Ganesan et al., 2013).

During mammographic interpretation, the radiologist combines the information from the two views with the assumption that if a mass appears in one view, most of the time the mass can be found as well on the other view (Blanks et al., 1999). This helps to identify the

object as a true or a false mass and to obtain a more accurate cancer detection by radiologists, in contrast with the separate evaluation of each view. Notice that, although the examination takes into account the two views, there are particular cases where a mass is only visible in one of the views.

Moreover, in recent years, the number of computer-aided detection (CAD) algorithms for mammography has increased rapidly, and its advantages and results confirm such increments on breast cancer detection (Brem et al., 2003). Nevertheless, the majority of CAD algorithms reported in the literature analyzed each breast view separately. Although, in general, high sensitivity of the CAD systems involves the presence of false positive (FP) detections.

This project aims to improve the performance of an already implemented framework for mass detection in mammograms by reducing its FP rate. The method proposed is based on the two-view image analysis, following radiologists' strategy when performing a diagnosis. The study is centered on a 3D breast deformation method which searchs for possible positions in the MLO view of a lesion detected in CC view. This is done by projecting epipolar curves in the MLO, which corresponds to the lesions in the CC view, and evaluating distance conditions regarding the MLO lesion to the epipolar curves to interpret the lesion as a mass or no mass.

At the same time, a pectoral muscle segmentation approach has been investigated as pre-processing step to discard potential lesions within this tissue Notice that this approach by itself can be considered as a FP reduction strategy.

2. State of the art

CAD systems aimed for breast cancer screening have been implemented since recent years, however, in the topic of detecting masses is still facing the complexity of the task. The main drawback of these methods is the high number of FP detections, which means that a CAD system interprets a normal tissue as a suspicious one. Approaches combining the information of the two views of the same breast have been proposed with this issue (Destounis et al., 2004). Previous research related to the general approach proposed in this project is described below:

2.1. Pectoral Muscle Segmentation

One of the main obstacles in mammographic image analysis is the presence of pectoral muscle (Ganesan et al., 2013). The extraction of this region has become a challenging task due to the issue that the density and texture information of both, pectoral muscle and breast tissue, are almost similar.

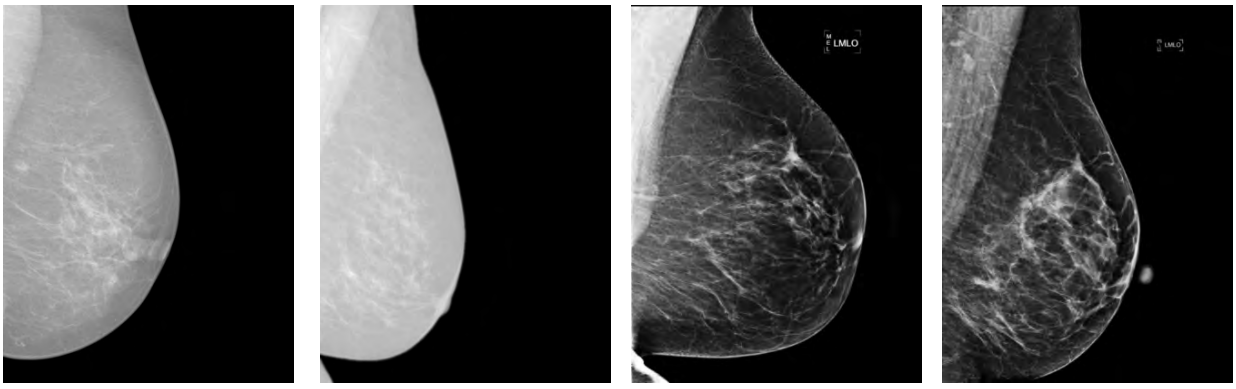
Therefore, pixels based segmentation techniques (Sultana et al. (2010), Saltanat et al. (2010) and Yapa and Harada (2008)) or techniques based on curvature of the edge of pectoral muscle (Ferrari et al. (2004) and Xu et al. (2007)) had a limited success in obtaining accurate results for a wide range of datasets. The study of Kwok et al. (2001) used a Hough transform to find the contour between breast tissues and the pectoral muscle, producing generally acceptable results.

Since recent years, the performance of convolutional neural networks (CNN) for tasks as segmenting, detecting and classifying objects on natural scenes has increased; thus, there was a keen interest to develop CAD systems applying this technique on the medical imaging area. That is the case of the recent approach of Rodriguez-Ruiz et al. (2018), where a method to segment the pectoral muscle in Digital Breast Tomosynthesis (DBT) using a deep learning approach is proposed. Their results show a recognized performance with images from different modalities (mammograms and synthetic images from breast tomosynthesis) with a median Dice Similarity Coefficient (DSC) value ranged from 0.947 to 0.977.

2.2. Correspondance views

The importance of analyzing the CC and MLO views by radiologists (ipsilaterality) has been reported since long time ago. For instance, the study of Warren et al. (1996) reports an increased in detection rates from 7.6 per thousand to 8.2 per thousand women screened, with 14 more cancers detected by radiologist when examining the two views rather than one view only.

CAD systems aimed at mass detection in mammograms with a two-views analysis approach are not as popular as the common examination and analysis of views independently. Wei et al. (2009) explored the implementation of fusion the information of the views CC and MLO with a dual CAD system, which merges the decision from two mass detection systems in parallel. The study was based on the identification of potential pairs of mass candidates by using a regional registration technique and applying similarity measures focused on the paired morphological feature. Then, each object was scored by computing Hessian and texture features for each mass pair identified, and finally, a discriminant analysis classifier between the 3 scores, the individual score per view and the one from the masses paired, was applied. This method obtained an average case-based sensitivity improvement from 67.4% to 83.7% for average masses, and 44.8% to 57.0% for subtle masses at the same FP rates.
(Hartley and Zisserman, 2003).



(a) (b) (c) (d)

Figure 1: Comparison of mammography images from different systems containing on the datasets. All images from INbreast dataset were obtained from a Siemens – *MammoNovation* (a). The Optimam dataset subset using on this project contained images obtained from several systems: General Electrics (GE) – *Senographe Essential* (b), Hologic – *Lorad Selenia* (c), and Hologic – Selenia Dimensions (d).

Kita et al. (2001) proposed a technique based on the calculation of curved epipolar lines by developing a 3D model of the deformation of the breast caused by compression in different view. Their approach was based on the assumption that it is possible to find the epipolar geometry relating pairs of images for further matches to a line. This means that a point in one image (e.g. CC view) defines a line in the other one (e.g. MLO view) on which the corresponding point must lie

Geometrical principles and deformable object behavior were the strategies conducted in that studies. Also by studying the guidelines for performing mammograms, they propose approximation of the breast behavior when being compressed. Some of the essential assumptions that they do are: The nipple keeps the same coordinates due to it hardly moves under compression; the compression force applied to the breast is always constant; the breast tissue deforms uniformly. The performance of the method was evaluated on a dataset of 37 lesions, and their method could predict the location in the second view by using a minimum distance tolerance of 6.78 ± 5.85 mm from the lesion in MLO to the curved epipolar line projected from the CC lesion.

3. Materials

3.1. ***Data sets***

Two datasets were used in this project, Figure 1 illustrates example of images containing in the datasets described below.

3.1.1. INbreast

The INbreast dataset is composed of full-field digital mammography (FFDM) images from screening, diagnostic and follow-up cases, acquired at the Breast Centre in the Hospital de Sao Jao, Porto, Portugal. Moreover, this dataset contains annotations of pectoral muscle and 6 types of findings: asymmetries, calcification, distortion, masses, multiple findings, cluster of micro-calcifications.

The dataset has a total of 115 cases, where 90 cases correspond to patients with 2 breasts, and the remaining 25 cases are from patients who had a mastectomy. In total, 410 images are contained on the dataset; 203 are CC views, 1 is caudio-craneal from below (FB) view and 206 MLO views. It is important to point out that from the 206 MLO images, only 201 have the present of pectoral muscle. Therefore, these 201 MLO images were used in this work to evaluate the pectoral muscle segmentation algorithm investigated.

Several characteristics of the dataset are listed below.

- Images were acquiered using the Mammo Novation Siemens system, see figure 1a.
- Images size of 3328x4084 or 2560x3328 pixels, according to the x-ray plate used, which depends on the breast size of the patient).
- Pixel size of 0.07 microns.
- Contrast resolution of 14-bits.

3.1.2. OMI-DB

The Optimam Mammography Image Database (OMI-DB) is an extensive mammography image database of over 80,000 unprocessed and processed digital images extracted from the National Breast Screening System (NBSS), which also contains expert-determined ground truths and associated data linked to the images.

From this database 54 cases with MLO images were selected. The number of the cases used on each Module (explained in follow section) is as follow: For the percoral muscle segmentation all 50 cases was taking into account, giving a total of 54 MLO images

were analysing. On the other hand, due to the random selection of the cases, not all of them were contained masses annotation findings. At the end, a total of 46 cases corresponded to patients with at least one lesion visible in both CC and MLO views. Each breast has its CC and MLO views, which gives a total of 92 images, as the dataset used to test the FP reduction scheme of this project.
Those images were obtained from different systems (Figure 1b, figure 1c and Figure 1d).

OMI-DB contains mammography images from several vendors and scanners. Each of which has different image characteristics.

1. General Electric (GE) vendor
   (a) Images obtained using a Sennographe scanner (Figure 1b).
   (b) Two types of mage size: 1914 x 2294 and 2394 x 3062.
   (c) Pixel size of 0.094 x 0.094 microns.
   (d) Contrast resolution of 12 bits.
2. Hologic vendor
   (a) Images obtained from two scanners: Lorad Selenia (Figure 1c) and Selenia Dimensios (Figure 1d.
   (b) Two types of image size for both scanners: 2560 x 3328 and 3328 x 4096.
   (c) Pixel size of 0.07 x 0.07 microns in the case of Lorad Selenia scanner. And 0.065 x 0.065 microns for the Selenia Dimensios one.

4. Methods

The primary objective of the project is to reduce FP detections as part of a framework for mass detection; the approach is focused mainly in the lesion correspondence analysis from the CC to the MLO view (ipsilateral analysis).

The complete workflow of the project is shown in Figure 2. Module A corresponds to the framework that is already implemented on our research group (i.e. VICOROB). Since the CAD system did not have any pre-processing for the pectoral muscle in MLO views, and it is required for module C, a pectoral muscle segmentation algorithm, using the deep learning approach of Rodriguez-Ruiz et al. (2018), was implemented in module B. Notice that this module by itself will be evaluated as a FP reduction too.

For the module C, due to the limit of time and resources, an already existing MATLAB code implementation of the method described on Kita et al. (2001) was adapted.

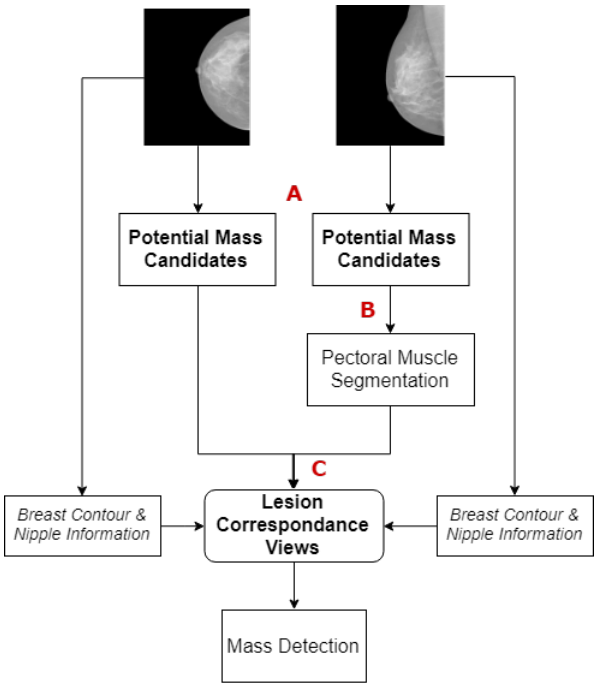


Figure 2: Workflow of the project divided into modules. Module A extraction of potential mass candidates per each view; Module B pectoral muscle algorithm for MLO views; and Module C the lesion correspondence between the two views.

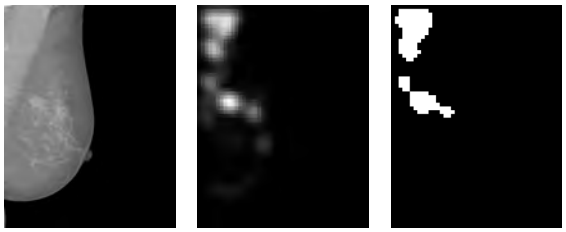
4.1. Potential Mass Candidates (Module A)

The potential mass candidates are generated by a framework for mass detection in mammograms developed in our research group. This framework is based on CNNs and makes use of 2 datasets to train the net, Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) and INbreast datasets. The CNN architecture used on this framework corresponds to a 50 layers of Residual Network (ResNet-50) previously pre-trained on natural images using the ImageNet dataset, and it is a path based fashion, with patches of size 224 x 224 pixels.

In summary, the framework works as follow. As the first step, it is wanted to transfer the domain of convolutional features from natural images to digitized mammograms. Therefore the net is trained with positive and negatives patches, generated from images of the CBIS-DDSM dataset, as follow:

- The net is trained only on the last fully connected layer by freezing the lowers layers.
- Then, the net is trained on all the layers.

Finally, the network is adapted to detect masses using path classification by fine-tuning the net using fully digital mammograms from the INbreast dataset. The output of the net is a pixel-wise probability map for breast masses.



(a) (b) (c)

Figure 3: Example of potential mass candidates extracted from INbreast dataset. (a) corresponds to the original image; (b) is the probability map (values from 0 to 1) resulted from Module A, and (c) displays the potential mass candidates obtained with a 0.70 theshold on (b).

For this framework, Free-Response Receiver Operating Characteristic (FROC) was used to evaluate it at different thresholds, in order to find an optimal configuration. Figure 3 shows an example of Module A output and the extraction of potential mass candidates from it.

4.1.1. Local threshold percentage

The extraction of the potential mass candidates are done by using a local thresholding to discriminated the 20% of the values contained in each probability map. This means that the value of the threshold is not fixed, instead, it change according to the maximum probability found on each probability map evaluating. Equation 1 illustrates the calculation of the local threshold.

$$Thresh = max(I) - \frac{max(I)\dot{T}_p}{100} \quad (1)$$

where I represents the probability map image, T_p is the percentage value wanted to restring, in this case is set to 20 and $Thresh$ is the local threshold value per probability map.

*4.2. **Pectoral Muscle Segmentation (Module B)***

Although the main objective of the project is not pectoral muscle segmentation, we is needed to have a robust algorithm for this matter. Therefore we implemented and adapted the algorithm described in (Rodriguez-Ruiz et al., 2018). This had to be adapted as the original work was applied to breast tomosynthesis images. Figure 4 illustrates the structure of the algorithm for segmenting the pectoral muscle. Each section of this pipeline (i.e. pre-processing, model prediction, and post-processing) are described below.

4.2.1. Pre-processing

When working with images that do not share the same characteristics, like the intensity ranges, size of the images, etc., it is essential to have a pre-processing stage to normalize and specify the data that enter into the net. This will help the network to learn the main

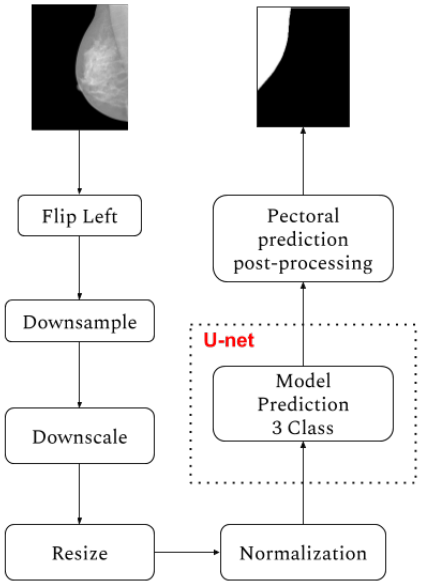


Figure 4: Structure of the Pectoral Segmentation approach.

characteristics of the image in a more generalized way. This is why a series of pre-processing and normalization were applied to all the images.The first step of the normalization is to corroborate that all the images passed to the net are MLO view. As the algorithm works with images with the standard Digital Imaging and Communications in Medicine (DICOM) format, it can discriminate CC views among the images by looking into the tag "ViewPosition" of the DICOM header.

Another important point is that all the MLO views must lay the breast tissue to the left part. Depending on the image type there are two options to flip the image. If the images are in DICOM format, the orientation of the image (i.e. right or left) can be known by means of reading the DICOM header. Otherwise, the algorithm compares the sum of the intensities of a 20 x 20 pixel region of the top left and right; the one with the maximum value indicates the position of the breast. Notice that, looking for a sumatory equal to zero on the opposite top side of the breast does not always work due to the variability of the intensity ranges as in some images the background is represented with low values different from zero.

After the image is flipped to the left, the image is downsampled to 1x1 mm^2 pixel size. If the image is with intensity values, a pixel area relation is applied to resampling; otherwise, if the image corresponds to the ground truth (binary image), a nearest-neighbor interpolation is used instead. Moreover, to handle the variability of the intensity ranges, all the images were downscale to 8 bits (0-255 grayscale).

At this point, the original size of the image has been reduced to a reasonable size for processing into a CNN. However, the sizes are not proportionally and equal among each other; therefore a zero padding/extraction method is used to homogenize the image size to 320 x 320 pixels, as the original proposal. This technique is affordable because, as all the images are flipped to the left and the breast tissue barely touches the right and lower image border. Hence, there is no possibility to alter the breast tissue. In addition, all the images were normalized by subtracting the dataset mean and divide it by its standard deviation. This serves to center the data and give more control stability to the net.

4.2.2. Prediction Model

For the prediction model, the U-net network as a 3 class model was used. The class corresponds to background, breast tissue & pectoral muscle. The U-net architecture follows the same level structure as the original one from Ronneberger et al. (2015), as despicted in Figure 5.

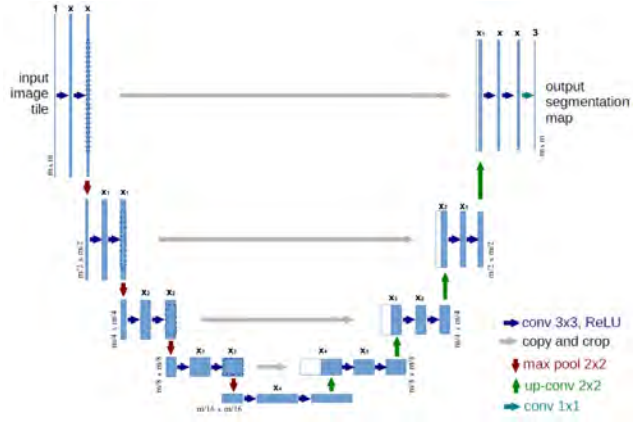


Figure 5: Structure of U-net architecture adapted from (Ronneberger et al., 2015).

As observed in figure 5, the number of filters for each convolutional layer is not specified because more than 1 model was implemented to exploit the behavior of the net. The description of the number of filters is explained on Section 5.The last layer of the net is a softmax activation function which gives a probability map of the pixels belonging to one of the three classes, with the same input size.

4.2.3. Post-processing

From the probability map of the 3 classes, it has to be extracted the class of interest, the pectoral muscle. Therefore an *arg max* function, which gives to the pixel the value of the class with the maximum probability, is applied. Figure 6 illustrates an example where, although the breast tissue and background classes are predicted well (reason of uniform colors), the hole's border inside the pectoral muscle shows a blurred color. Examples were is visualize better this behavior can be found on Experiment Section. Notice that figure 6b is displayed in colors although the probabilities go from 0 to 1 as the example of figure 3b. As each pixel has a probability for each of the three class, the image can be interpreted as an RGB image due to it has 3 channels:

- First channel: Background class probability (Red color).
- Second channel: Pectoral Muscle class probability (Green color).
- Three channel: Breast tissue probability (Blue color).

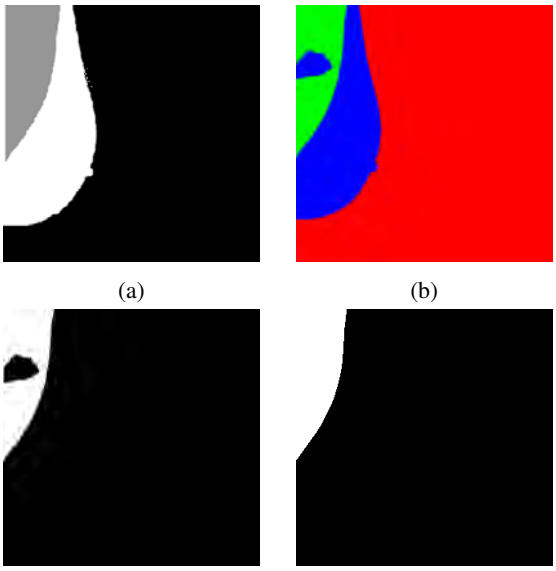


Figure 6: Example of post-processing steps followed. (a) three class ground truth. (b) three class probability map resulted from the net prediction where. (c) pectoral muscle class extraction. (d) fill-holes algorithm. The label of the images are: Red and Black = background, blue and white = breast tissue, and green and gray = pectoral muscle.

After that, the prediction class extracted is upsampled to the original resolution using linear interpolation. Since segmentation can fail during the process, for instance, incomplete or over-segmentation, a fill-holes algorithm is applied. This algotihm first look for the number of object existed on the prediction image by applying a connected components and find contourts functions, if there are more than one object, the algorithm selects the bigger one closer to the left upper border. Then, the algorithm fill the object.

A sample of segmentation failure (i.e. hole)is shown in Figure 6c and the result after applying the fill-holes algorithm, Figure 6d. Notice that this algorithm also take cares of segmentation failures when more than one object is predicted as pectoral mask.

4.2.4. Evaluation

Each pectoral muscle prediction was evaluated using the Dice Similarity Coefficient (DSC) with the Ground Truth, see Eq.2. 2.

$$DSC = \frac{2|Pect^{Model} \cap Pect^{Truth}|}{|Pect^{Model}| + |Pect^{Truth}|} \quad (2)$$

where $Pect^{Model}$ refers to the pectoral prediction mask, and $Pect^{Truth}$ corresponds to the pectoral muscle ground truth.

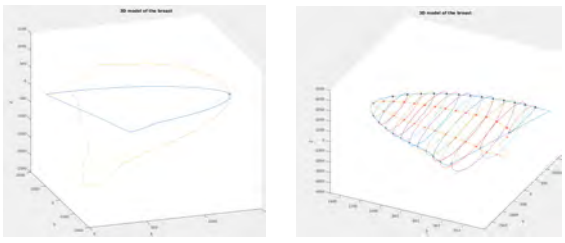
*4.3. **Lesion correspondence on CC & MLO Views***

As mentioned before, the two-view lesion correspondence part was based on the Kita et al. (2001) method, and an existing MATLAB code was adapted. Notice that, in order to compute a 3D breast deformation, there are certain information required:

1. Acquisition information, usually found in the DICOM header: pixel spacing, the distance between the x-ray and the compression paddle, and the breast thickness under compression.
2. Coordinates of the breast profile and nipple.
3. MLO views with pectoral muscles segmented.

Point 1 is needed to obtain by mathematical and geometric forms the decompression and compression information of the CC and MLO view respectively, which are used to obtain the behavior of the breast deformation.

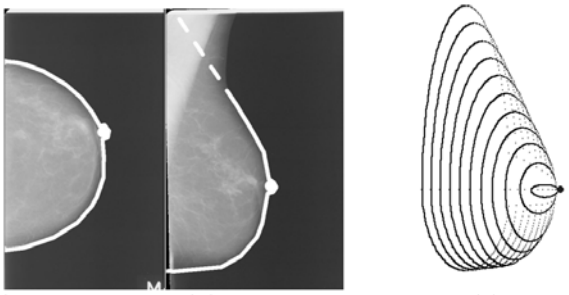
Point 2 & 3 are aimed to construct the 3D model of the breast. The reason why the pectoral muscle segmentation plays an important role here is that, when approximating a 3D model of the breast, the pectoral muscle is not present in CC view, so the 3D model will not work correctly.



(a) (b)

Figure 7: 3D model computation. (a) represents the aligment of the MLO and CC contours based on nipple position. (b) corresponds to the 3D breast model from (a).

Notice that in the implementation of Kita et al. (2001), the nipple coordinates and the breast outlines are extracted manually. Also, for computing the breast 3D model they approximate the outline of the MLO with the contour of the Medio Lateral (ML) or Latero Medial views because it was available, Figure 8 represents an example of a 3D reconstruction.



(a) CC image (b) MLO image (c)

Figure 8: Illustration of 3D reconstruction of the breast approach implemented on Kita et al. (2001).

In contrast, for the approach proposed here, the following steps were made:

- For Point 2 previously mentioned only the nipple coordinates were taken manually. For the breast profile, a simple threshold and connected component approach was implemented.
- To deal with the pectoral muscle presence, the pectoral segmentation algorithm describded in section 4.2 was used.

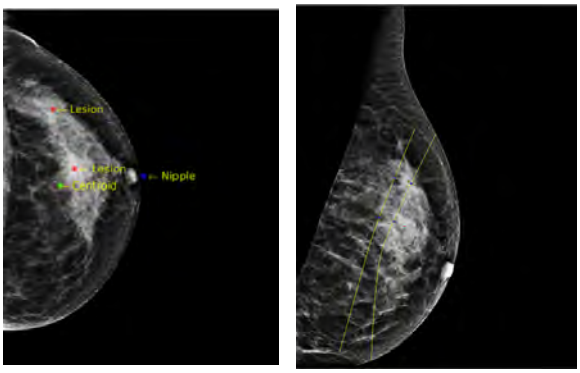
The existed MATLAB code was oriented for a medical area, aimed for use as a visual tool for radiologist. The code assumes MLO images have already segmented the pectoral muscle, and that both images, CC and MLO had masked the breast to delete external artifacts. For the case of the nipple and lesion coordinates, they were obtained manually by the user in real time the program was analyzing the correspondence.

For this project it was needed to modify and automatize the mentioned MATLAB code. For instance, the nipple coordinates were previously taken and stored to later be accessible to the algorithm. Furthermore, apart of working with the intensity image, the probability map obtained from Module A is used as well. Then, for the breast profile and pectoral masking is calculated and applied, respectively, during the computation of the algorithm. Excluding the modifications previously mentioned, the MATLAB algorithm performs the same steps described in the original method. A brief summary is explained below:

First, the algorithm adjusts the CC and MLO view. The CC view must have the centroid of the breast aligned with the nipple coordinate. On the other hand, for the MLO view, the aligment of the centroid and the nipple have to be with a 45 angle. Once the views are adjusted, they are aligned within the nipples

coordinates to build the 3D model. Figure 7a displays the alignments of the views, the CC contour lays on the X axis, while the MLO sits on the Y axis. Figure 7b shows the following step, the 3D model computation.

After that, with the technical information of the CC and MLO compression, the epipolar lines can be calculated for present masses on CC view by applying decompresion algorithm to the centroid of the mass; and further projecting the lines on the MLO view using an uncompression algorithm, which transform the epipolar line into an epipolar curve. A more detail information and mathematical explanation can be found on the original propose of (Kita et al., 2001).



(a) (b)

Figure 9: Lesion correspondance view example. (a) corresponds to the CC view, red points are the lesion detected, green poit is the centroid of the breast, and blue one is the nipple coordinates. (b) illustrates the MLO view of the same breast with the curves projected from lesion in (a).

Figure 9 illustrates an example of epipolar curves in the MLO projection from 2 masses detected in the CC view.

4.4. False Positive Reduction

The FP reduction strategy consist in make use of the epipolar curves projected on the MLO view, obtained from section 4.3. Basically, an Euclidean distance calculation (Eq. 3) is applied from the centroids of MLO candidates masses to the epipolar curve(s) projected.

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{3}$$

where x_1 and $y1$ corresponds to x and y coordinates of one point; and x_2 and $y2$ to a second point.

The distances measured are subject to followed conditions to determinate the interpretation of each mass candidate.

1. If one candidate mass on the MLO lies close to an epipolar curve projection of a CC mass candidate, and the distance between is less than 20mm, both candidate masses are considered as a correspondence lesion. Thus, the masses on both views (CC and MLO) are classified as true positive detection.
2. Otherwise, if the distance to the epipolar curve is too large or there is no object in correspondence, the candidates masses of each view are discarded and interpreted as a false positive detection.

4.5. Computational Environment

The computations were performed on a Linex workstation, Ubuntu 18.04, with 12 CPU cores and a single NVIDIA GeForce GTX GPU with 125,6GB memory.

The deep learning framework used is Keras-2 with Tensorflow as backend using Python 2.7 environment. Regarding the two-view lesion correspondence, MATLAB R2016a program was used.

5. Experiments

5.1. Pectoral Muscle Segmentation

The first experiments for this module were made using the INbreast dataset due to the existence of its pectoral muscle ground truths. The images of this data were split into training, validation and testing sets with a proportion of 60%, 20%, and 20%, respectively. The three sets were balanced using the number of cases per each density category. This means the sets contains equal difficulty levels of cases and none of the images from the same patient belongs to more than 1 set. Furthermore, the OMI-DB sub-dataset was used for testing the net as it contains images from different systems and with different intensity values distribution, see Figure 1. Table 1 shows the distribution of the images for each dataset. For the experiments described below were used only INbreast images. The images from Optimam were only for evaluation.

Table 1: Distribution of images from different datasets for training, validation and testing the models. The number of system refers to the scanners used to obtain the images, explained on section 3.1.

Distribution of datasets				
Dataset	**Set**	**System**	**Patients**	**Images**
INbreast	Training	1	66	120
	Validation	1	17	32
	Test	1	20	36
Optimam	Test	2	8	8
		3	12	12
		4	14	14

5.1.1. U-net configuration

Before the computational system mentioned before was available for this project, the first u-net configuration for pectoral muscle segmentation was tested on a Linux workstation, version Ubuntu 16.04, with 8 CPU cores and one NVIDIA GeForce GTX GPU with 12GB memory. In addition, different architectures configuration were implemented. Table 2 describe them.

Table 2: Description of number of filters per each convolution layer of the 3 U-net models used in this work.

U-net Model's Configurations

Model	x	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$
U-net 1	16	16	32	32	64
U-net 2	32	64	128	256	512
U-net 3	64	128	256	512	1024

5.1.2. Selection of post-processing strategy

For the post-processing of the pectoral prediction image was experiment the fill-holes algorithm, described on section 4.2.3, with and without binary dilation to the image, Figure 13 illustrates an example. Futhermore, an evaluation of applying them before or after the image is upsampled to its original resolution was done as well.Note that Figure 13 only compares the aspect of the pectoral border rather than the segmentation precision. Therefore the ground truth of the image is not displayed.

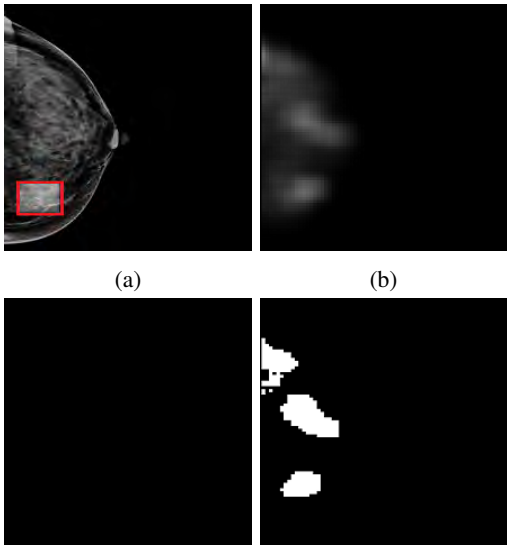
(a) (b)

Figure 10: Comparison of upsample without dilation (a) and when using it (b). Dilation of 3x3 window. Notice that the image has been cropped for visual displayment.

*5.2. **Potential Mass Candidates***

As explained before on section 4.1, the output of Module A is a probability map with a 0 to 1 range values. The potential mass candidates are extracted by applying a threshold to the probability image. The first technique applied was a fixed threshold value for all the images (global threshold), however, zero detections of mass candidates were faced; caused when the image have in general a low probability map.

Thus, a local threshold method was opted to avoid modifying the probabilities by normalizing all from 0 to 1. This method looks for the maximum value of each probability map and calculate the threshold values according to the percent of probability is wanted to take into account (local threshold). A high percent value, i.e. 90-80% means a high threshold, and inverse with low percent values.



(a) (b)

(c) (d)

Figure 11: Global threshold and local threshold comparison for potential mass candidates extraction. (a) Original intensity image, (b) probability map (from 0 to 1), potential mass candidates at global threshold of 0.5 (c), and at local threshold of 50% (equal to 0.21) (d).

5.2.1. Local threshold percentage

Figure 11 shows a clear example of comparing the output of a global threshold against a local one. Figure 11b corresponds to the probability map, which has 0.43 as maximum probability value. When applying the global threshold of 50% over 0 to 1, which is equal to 0.5, the results are zero detections, figure 11c.

In comparison, the use of a local threshold will always detect at least one candidate mass, Figure 11d, because, in this case, the threshold value instead of being 0.5, change to 0.21, as the maximum probability of the image is 0.43. Equation 1 illustrates the calculation of the local threshold.

Previously, the example made for the local threshold percentage was using a 50%. However, 9 values were experiment in orden to select and fix the percentage of restrinction. Therefore, the three modules, A, B and C, were subject under different thresholds percentage values to compare and select the one were the algorithms performs better.

Table 3: Modules performance comparison under different thresholds

View	Module	Class	GT	Thresholds (%								
[0.5ex]				10	20	30	40	50	60	70	80	90
CC	A	TP	26	16	18	20	23	21	20	21	14	3
		FP	0	12	17	25	27	38	39	50	68	56
		FN	0	8	6	5	4	4	0	0	10	21
	C	TP	26	8	10	9	10	10	11	11	9	2
		FP	0	5	5	11	13	17	17	17	25	9
		FN	0	18	16	17	16	16	15	15	17	16
MLO	A	TP	26	16	18	20	23	21	20	21	14	3
		FP	0	20	28	39	40	56	77	98	125	106
		FN	0	10	8	6	3	5	6	5	12	23
	B	TP	26	18	21	23	24	24	26	17	12	3
		FP	0	15	22	32	42	61	83	79	107	113
		FN	0	8	5	3	2	2	0	9	14	23
	C	TP	26	10	8	8	10	6	6	2	2	0
		FP	0	5	9	14	15	19	16	26	27	10
		FN	0	16	18	21	20	20	19	23	22	18

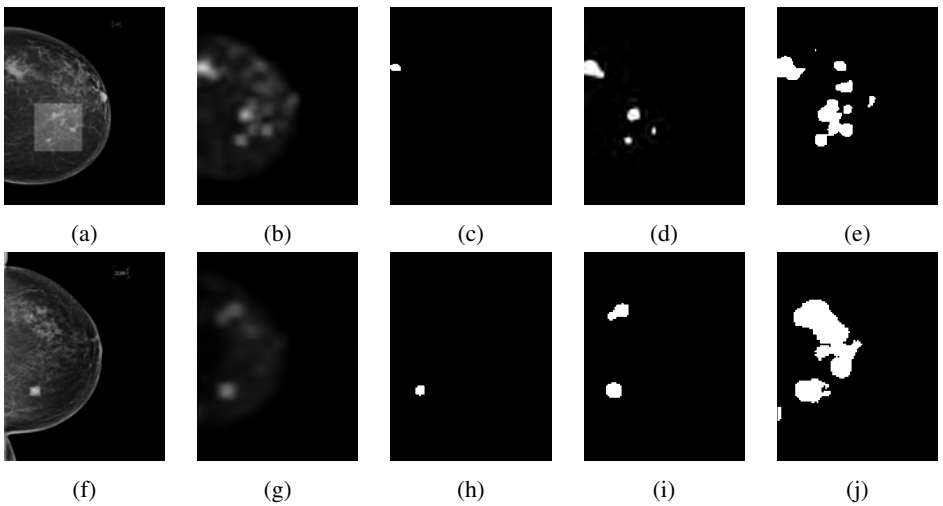


Figure 12: Comparison of potential mass candidates extraction with local threshold. (a) & (f): intensity images. (b) & (g): probability maps. (c) & (h), (d) & (i) and (e) & (j): potential mass candidates at 20%, 50% and 80% threshold, respectively.

Table 3 reports the result of each module per view in terms of classifying the masses as TP, FP and False Negative (FN) classes. Notice that, for the evaluation of the module C in this case, the distance tolerance to classify the mass candidates as mass or no mass was 100 pixels.

Figure 12 illustrated the behaviour of the potential mass candidates extraction within different threshods. Therefore, the final experiments were decided to be conducted with a threshold of 20%.

6. Results

6.1. Pectoral Muscle Segmentation

Therefore, the first configuration on the model was "Unet-1". However, even with a significant reduction regarding the number of filters per convolutional layer the network was given a reasonable performance with this configuration; this can be possible due to the positional information to the net (the pectoral muscle always lay on the left upper part of the image) and the 3 class approach. After improving the computational resources, the following model's configurations, U-net 2 & U-net 3 were tested. The first pectoral muscle segmentation results for both models were better, in comparison of the first configuration. All the results show in this section use a guide a red line representing the ground truth contour, and the yellow line the pectoral prediction contour. Figure 13 displays example result of the three models output.

The table 4 shows the result of DSC evaluation over the 2 datasets. As the OMI-DB is integrated with images obatained from different systems it was decided to present the results for each one separately. As can be noticed, the performance of the net when evaluating images from the same system is considarable. see Figure **??**

Dataset	System	Images	DSC
INbreast	1	20	0.94
Optimam	2	8	0.72
	3	12	0.80
	4	14	0.79

Table 4: Distribution of images from different datasets for training, validation and testing the models. The number of system refers to the scanners used to obtain the images, explained on section 3.1.

The net fails in a few cases, as the Figure 15 shows, when the contour of the pectoral muscle is not well defined or have a similar structure as the breast tissue.

On the other hand, when testing images coming from another system, the net is not able to generalize due to the distribution of the intensities are quite different from the one the net learned.

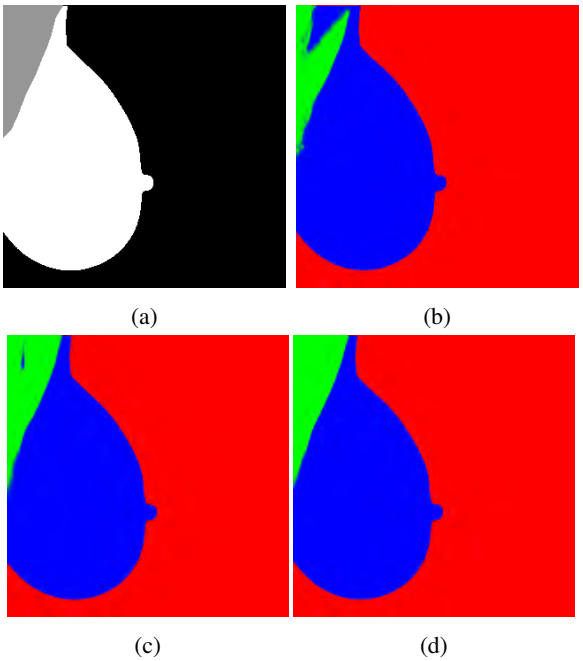


Figure 13: U-net prediction comparison using different number of filters for the same image. (a) is the three class ground truth. (b)-(d) display the three class probability map predictions from configuration models *'U-net 1'*, *'U-net 2'* and *'U-net 3'*, respectively. Its configurations are found on Table 2. The label of the images are: Red and Black = background, blue and white = breast tissue, and green and gray = pectoral muscle.

The Figure 16 illustrate the highest DSC obtaining from each system from Optimam dataset.

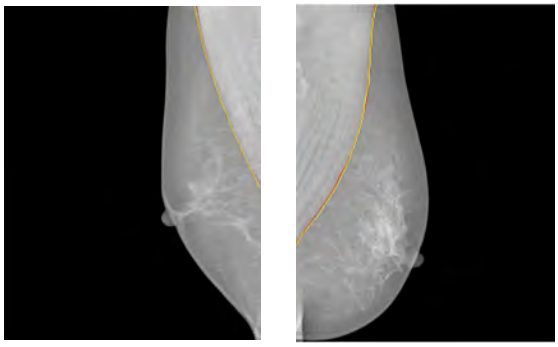
And finally, below are displayed the lower DSC from those three system.

6.2. False Positive Reduction

As the performance of the pectoral muscle segmentation on the Optimam data was not satisfactory in most of the cases, it was decided to evaluate the performance of the FP reduction without the interference of the pectoral muscle algorithm's error. This means that the groung truth of the pectoral muscle was used instead.

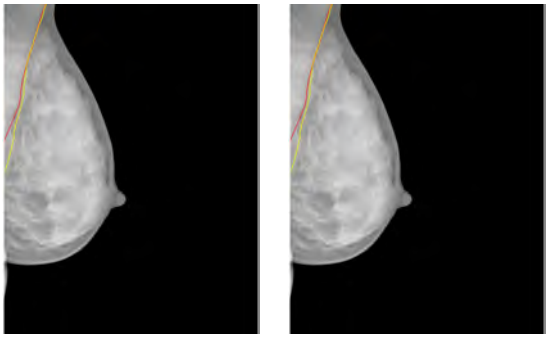
The Table 5 contains the results of all the models. As first look, incongruence within resulst of MLO with and without pectoral muscle segmentation come into the mind. Logically, the TP should not be alterated, at least that one mass lays on the pectoral muscle; and the FP should be reduced; masses from pectoral muscle are discarded not added. However, this logical can not be applied the approach proposed because:

1. The pectoral mask is applied to the image directly, without any area or region analysis. Therefore, if one uniform mass is between the pectoral and the



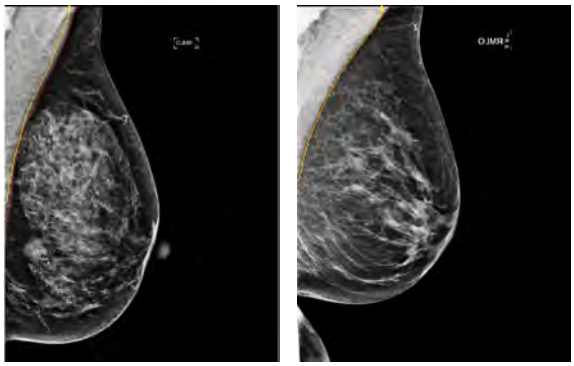
(a) Ground Truth (b) U-net 1

Figure 14: Example of pectoral muscle segmentation result from INbreast images with DSC = 0.99



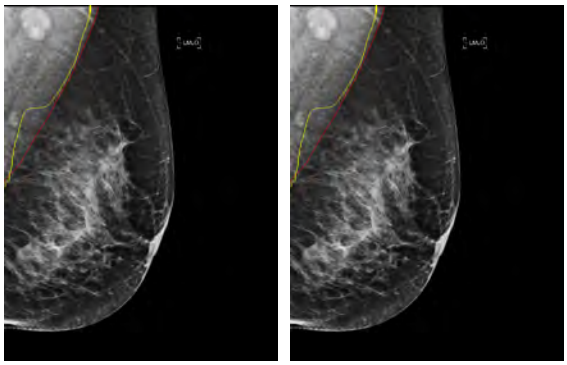
(a) (b)

Figure 15: Example of pectoral muscle segmentation result from INbreast images with DSC 0.96.



(a) (b)

Figure 16: Example of pectoral muscle segmentation result from Optimam images with DSC = 0.98



(a) U-net 2 (b) U-net 3

Figure 17: Example of pectoral muscle segmentation result from Optimam images with DSC = 0.85

breast tissue, instead of being one object, after applying the mask can partitionate it into several objects, see Figure 18a

2. As the evaluation of the mass takes into account the centroids, the calculation of this position will be affected to objects layed on the middle of the muscle.
3. And finally, an important factor of this behaviour is due to the pectoral is masked on the probability map image; this mean that, the calculation of the image threshold, according to its minimum and maximum probability change. For example, on Figure 18b, the blue contours were masses candidates detected on the pectoral muscle, however the high intensities of the image belongs to those masses. Therefore, when calculating the threshold, the intensity of the real mass, red contour, is not taken into account due to the low probability. When the probability map is segmented, the high probabilities now are moved to the real mass. Thus, a TP detection is summed up thanks to the pectoral segmentation removal.

Regarding the two-view lesion correspondence, there were some cases were the algorithm predict almost exactly the position of the lesion on the MLO, as show in Figure **??**. The minimum distance from the lesion centroid to the epipolar line was 0.2mm.

Most of the time were the program fails is caused since the mass candidates extraction. That is the case of Figure 20 were, visually, the epipolar line shows a clear correspondence between the two view, but the real mass was not taken into account as a mass candidate due to its low probability predicted from Module A.

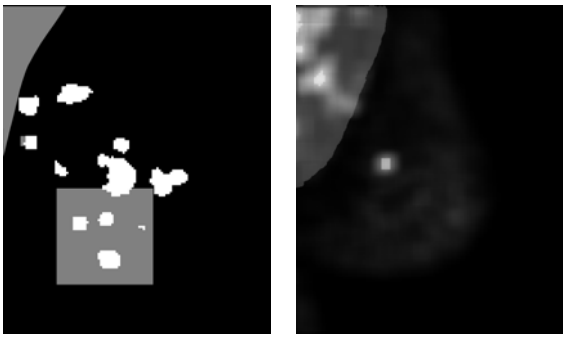
7. Discussion

In this section the results of the algorithms implemented are discussed and the conclusion of this work are given, in terms of single and full FP reduction strategy.

7.1. Pectoral muscle segmentation algorithm

7.1.1. U-net configuration

Three models with the same architecture but varying the number of filters per convolutional layer were implemented. The figure 14a is a clear example of the behavior of each model.
When a CNN has limited number of filters, the net is



(a) (b)

Figure 18: Example of aumenting FP findings (a) and TP (b)

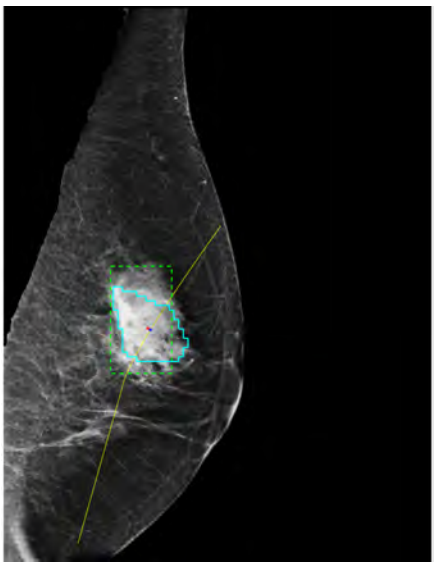


Figure 19: Example of good match using correspondence approach

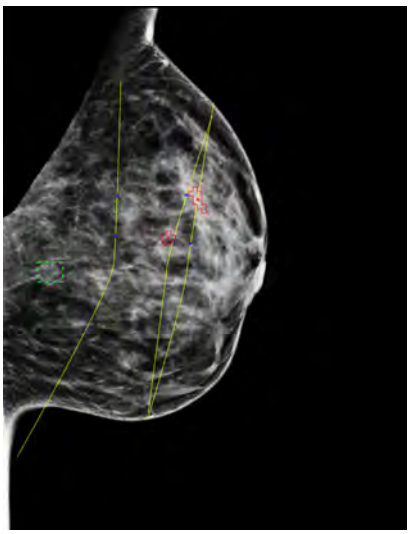


Figure 20: Example of bad result using correspondence approach

Table 5: Modules performance comparison under different thresholds

View	Module	Class	GT	10mm	15mm	20mm
CC	A	TP	26	20	20	20
		FP	0	17	17	17
		FN	0	6	6	6
	C	TP	26	10	12	15
		FP	0	5	6	8
		FN	0	16	14	11
MLO	A	TP	26	18	18	18
		FP	0	28	28	28
		FN	0	8	8	8
	B	TP	26	21	21	21
		FP	0	22	22	22
		FN	0	5	5	5
	C	TP	26	8	11	12
		FP	0	9	7	11
		FN	0	18	15	14

not able to generalize and interpret the image correctly, as noticed on the figure 14b. Therefore when the number of filters increase, the net can capture the essential characteristics of the problem evaluated, figures 15a- 15b, where the last one corresponds to the original structure described on Rodriguez-Ruiz et al. (2018).

7.1.2. Pectoral segmentation result

At the beginning of the project, it was decided to work only with one dataset, INbreast, due to the advantage of the pectoral muscle annotation. However, when the first test was performed on the CC and MLO correspondence algorithm, it was noticed that INbreast dataset lack of crucial information on the DICOM header due to data anonymization.

Due to the time limitation of the project, it was not possible to create manually ground truths of the pectoral muscle for all images to train the net for that dataset. Therefore, as the model trained and tested on INbreast dataset gives acceptable results for pectoral muscle segmentation, it was decided to test the model with another dataset, OMI-DB.

In order to evaluate the segmentation produced by the model for the OMI-DB, 46 pectoral muscle ground truth were manually annotated by two engineers with large experience in mammography. The annotations were computed using the ITK-SNAP software.

7.1.3. Post-processing strategy

For the post-processing was decided to include the fill-holes algorithm with the dilation step after the upsample of the image. because if the post-processing algorithm are applied before the upsampling, the image end-up with pixeled border. However, the border

appearance looks better when the post-processing algorithm are applied after. In this point, the dilation steps plays an important role for visual assignment, therefore, although this step no dot impact significantly the DCS calculation, it was decided to maintained on the pipeline.

Figure 10 shows a comparison of applying that the dilation do not have an important impact to the performance of the net, in comparison with the fill-holes scheme. and the slightly differents regarding the shape of the pectoral border can be notice.

7.1.4. Drawbacks

One of the commom issue found when working with medical images is the variability in the image characteritics. This is caused by two main subjects.

1. The intensity variation when taking mammograms cannot be controlled as it is subject to the automatic x-ray exposure control, which depends on the composition and internal distribution of each breast.
2. There is not an standard image processing technique that all the systems must followe when processing the RAW image. Therefore, each system applies different criterias depending on each vendor.

When the pectoral muscle segmentation algorithm was tested with images coming from different systems, this issue was faced. Thus, a normalization step for the images is needed. Possible approaches include background correction, or histogram matching.

Other alternatives to deal with this issue is training the net with a dataset containing images from different systems, then the net should be able to generalize for all the cases. In addition, avoiding large background areas on the images by cropping them within the breast tissue could help to give less weight to the background and focus on discrepancies between breast tissue and pectoral muscle.

On the other hand, using this approach as a FP reduction strategy, it gives good results in general when test it after module A. However, if the pectoral muscle is marked before producing the probability maps, its performance is expected to improve.

It is important to mention that, in particular cases, segmenting the pectoral muscle can compromise the diagnosis, since the presence of certain structures (e.g. abnormal axillary lymphs) can be useful to radiologists.

7.2. Potential Mass Candidates

The results show a better TP detections within middle local threshold in the two views. However, in these values, the potential mass candidates increase but behind the local threshold percent values can exist inadequate clinical tolerance. In other words, taking as example the local threshold of 40% and the example of potential mass candidates extraction of Figure 11, the probabilities taken into account for this case are over 0.17 probability values. It is true that only with this low value the mass can be detected, but clinically, classify a mass with less than 20 percent of probability of being a mass can not be allowed. Therefore the final decision was taking as a local threshold of 20% percentage to let pass the probabilities.

7.3. False positive reduction algorithm

The general performance of the algorithm proposed cannot be evaluated strictly in this project because is not possible to confirm the fails were caused by the algorithm itself. More experiments regarding the distance tolerance could lead to increase its performance but without reaching a significant grow.

The main issue is that this algorithm proposed is aimed to reduce FP detection. Some of the cases study in this project were having, in the Module A, a maximum probability less than 0.5 of being a lesion. Thus, when computing the potential candidate mass the TP rate was low. This was caused due to the module A was trained as well with the INbreast dataset. Therefore, when getting as input new images from other systems, the framework fails. The same problem that Module B presents.

Therefore, until the module A does not generate valuable potential masses that detects the majority or all the TP, the algorithm cannot exploit its potential. One solution can be the application of Module B, when normalization is done for working with images from different systems, before the MLO image enters into the framework of lesion detections.

8. Conclusions

Calculating the position of one object from one projection to another one, adding a deformable environment due to compression factor, has been a challenging task. The results expected were not possible to reach. The main drawbacks was the influence of external factors to the algorithm. For instance, breast profile containing none breast tissue, nipple detection precision, and the main dependency to the performance of an existed framework.

The belief that the idea of the project can work is not discard. However, the cost and benefits can discourage. If the algorithm can have an acceptable performance when the external factor, mentioned before, works correctly, then the algorithm comes into a second plane until those factors have an acceptable performance.

References

Blanks, R., Wallis, M., Given-Wilson, R., 1999. Observer variability in cancer detection during routine repeat (incident) mammographic screening in a study of two versus one view mammography. Journal of Medical Screening 6, 152–158.

Brem, R.F., Baum, J., Lechner, M., Kaplan, S., Souders, S., Naul, L.G., Hoffmeister, J., 2003. Improvement in sensitivity of screening mammography with computer-aided detection: a multiinstitutional trial. American Journal of Roentgenology 181, 687–693.

Destounis, S.V., DiNitto, P., Logan-Young, W., Bonaccio, E., Zuley, M.L., Willison, K.M., 2004. Can computer-aided detection with double reading of screening mammograms help decrease the false-negative rate? initial experience. Radiology 232, 578–584.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D.M., Forman, D., Bray, F., 2015. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. International journal of cancer 136.

Ferrari, R.J., Rangayyan, R.M., Desautels, J.L., Borges, R., Frere, A.F., 2004. Automatic identification of the pectoral muscle in mammograms. IEEE transactions on medical imaging 23, 232–245.

Ganesan, K., Acharya, U.R., Chua, K.C., Min, L.C., Abraham, K.T., 2013. Pectoral muscle segmentation: a review. Computer methods and programs in biomedicine 110, 48–57.

Hartley, R., Zisserman, A., 2003. Multiple view geometry in computer vision. Cambridge university press.

Kita, Y., Highnam, R., Brady, M., 2001. Correspondence between different view breast x rays using curved epipolar lines. Computer Vision and Image Understanding 83, 38–56.

Kwok, S., Chandrasekhar, R., Attikiouzel, Y., 2001. Automatic pectoral muscle segmentation on mammograms by straight line estimation and cliff detection, in: Intelligent Information Systems Conference, The Seventh Australian and New Zealand 2001, IEEE. pp. 67–72.

Misra, S., Solomon, N.L., Moffat, F.L., Koniaris, L.G., 2010. Screening criteria for breast cancer. Advances in surgery 44, 87–100.

Rodriguez-Ruiz, A., Teuwen, J., Chung, K., Karssemeijer, N., Chevalier, M., Gubern-Merida, A., Sechopoulos, I., 2018. Pectoral muscle segmentation in breast tomosynthesis with deep learning, in: Medical Imaging 2018: Computer-Aided Diagnosis, International Society for Optics and Photonics. p. 105752J.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241. doi:10.1007/978-3-319-24574-4_28.

Saltanat, N., Hossain, M.A., Alam, M.S., 2010. An efficient pixel value based mapping scheme to delineate pectoral muscle from mammograms, in: Bio-Inspired Computing: Theories and Applications (BIC-TA), 2010 IEEE Fifth International Conference on, IEEE. pp. 1510–1517.

Siegel, R., Ma, J., Zou, Z., Jemal, A., 2014. Cancer statistics, 2014. CA: a cancer journal for clinicians 64, 9–29.

Society, A.C., 2013. Cancer facts and figures 2013.

Sultana, A., Ciuc, M., Strungaru, R., 2010. Detection of pectoral muscle in mammograms using a mean-shift segmentation approach, in: Communications (COMM), 2010 8th International Conference on, IEEE. pp. 165–168. doi:10.1109/ICCOMM.2010.5509003.

Warren, R.M., Duffy, S., Bashir, S., 1996. The value of the second view in screening mammography. The British journal of radiology 69, 105–108.

Wei, J., Chan, H.P., Sahiner, B., Zhou, C., Hadjiiski, L.M., Roubidoux, M.A., Helvie, M.A., 2009. Computer-aided detection of breast masses on mammograms: Dual system approach with two-view analysis. Medical physics 36, 4451–4460.

Xu, W., Li, L., Liu, W., 2007. A novel pectoral muscle segmentation algorithm based on polyline fitting and elastic thread approaching, in: Bioinformatics and Biomedical Engineering, 2007. ICBBE 2007. The 1st International Conference on, IEEE. pp. 837–840.

Yapa, R.D., Harada, K., 2008. Connected component labeling algorithms for gray-scale images and evaluation of performance using digital mammograms. International Journal of Computer Science and Network Security 8, 33–41.

Medical Imaging and Applications

Master Thesis, June 2018

A fully automated deep learning quality assessment framework for online brain MRI processing

Roberto Paolella, Sergi Valverde, Arnau Oliver, Xavier Lladó

Computer Vision and Robotics Group, University of Girona, Catalonia, Spain

Abstract

Currently, most users manually select and analyse data through visual inspection, which is intuitive on a small scale but becomes impractical and error-prone on a large dataset. An entire research experiment may be distorted because of some no good quality images over all the data. Magnetic Resonance Imaging (MRI) can be corrupted from different preprocess algorithms. For instance, the skull stripping procedure commonly called brain extraction is often the first component in neuroimage pipelines and therefore, its robustness is critical for the overall performance of the system. Many methods have been proposed in the literature to address this problem. Moreover, there are many other processes where the images can be accidentally corrupted and none of these modules have a quality control of the obtained results. Hence, having an index that tells us about the quality of the images in an unsupervised way is essential for having good results in further phases. To ensure the quality of the acquired images we have developed a modular framework in which we deal with the different problems related to the images corruption after the preprocessing algorithms. In this master thesis three main process for plane recognition, brain extraction recognition and quality control of the skull-stripped brain were developed. We tackle these problems by combining some of the multiple modules developed in order to build different pipelines. One of these pipelines, for example, is designed to obtain the best parameter for the skull stripping tools. This pipeline as well the different modules have been successfully validated in different MRI brain images. Thanks to the good results obtained, the different modules implemented can be used in the future for an automated quality assessment of brain MRI scans.

Keywords: Quality assessment, preprocessing, framework, MRI, pipelines.

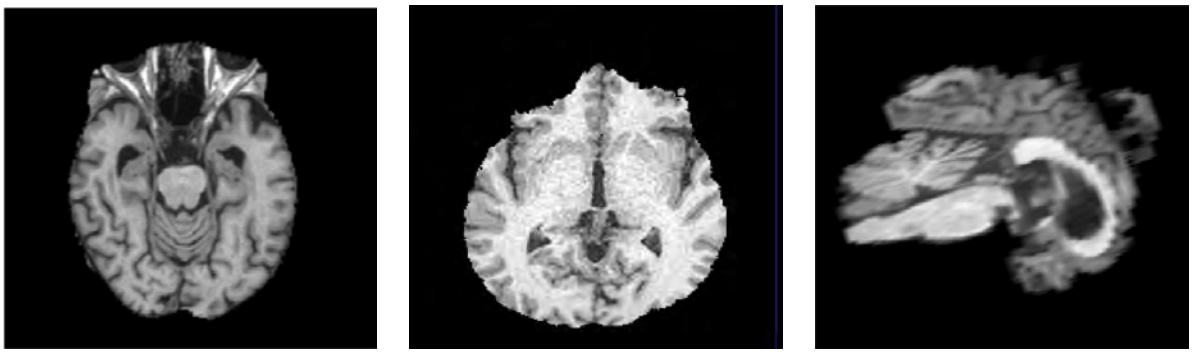
1. Introduction

Magnetic Resonance Imaging (MRI) has evolved into an essential diagnostic technique in medical imaging (Albers et al., 2006; Cerqueira et al., 2002; Warach et al., 1996). As a consequence of this, in the last few years, the use of MRI has grown exponentially. Furthermore, the different unsolved problems present in the medical imaging field such as longitudinal evaluation of the lesion, quantitative analysis of the different structures of the brain, etc. has lead to increasing collaboration between engineers and medical experts in order to achieve better solutions.

Due to the difference on the acquired MRI images used in the different medical algorithms, preprocessing plays an important role. Moreover, preprocessing algorithms are one of the preliminary steps that are required to obtain high accuracy on further steps. However, after applying some preprocessing techniques on the raw data, such as skull stripping, registration, denoising, bias correction etc, a wide variety of artefacts can degrade the reliability of the MRI images quality (see Figure 1). Image artefacts can compromise the utility of MRI volumes in brain studies. These artefacts can include a wide range of errors such as the one listed below:

- Incorrect brain extraction leading to unexpected structures such as eyes, part of the skull or others parts not considered as a brain.

- Structure of the volumes uploaded in the computer's memory: it is not always possible to rely on the header of the file for correctly interpreting the raw data in the software used. For example,



(a) (b) (c)

Figure 1: Example of three different bad quality MRI images. a) Inaccurate skull stripping. Presence of the eyes. The eyes must not be included when a skull stripping is performed. b) Excessive skull stripping. In this image is possible to see that the fundamental structure of the brain damaged. c) Orientation error. The orientation of this image has been corrupted after a registration process.

it is crucial to recognise how the different planes (sagittal, axial, coronal. see Figure 2) are stored in the uploaded volume. This problem is also known as the individualisation of the axis where the views are stored in the matrix[x, y, z].

- Quality control of the registration process.

The idea, in this master thesis, is to build a dynamic framework where it will be possible to add solutions for all the problems related to the Quality Control (QC). Solving the mentioned problems is important because failure in recognising one of these artefacts can cause various errors in the future morphometric analysis. Due to the possible involvement of the structure of the brain, the propagation of these errors into subsequent analysis can lead to incorrect diagnosis results. Nowadays the method most often used to avoid the dissemination of error along image analysis pipelines is the use of visual QC verification step. This step is performed manually by experts. The downstream image assessment thus becomes susceptible to intra- and inter-evaluator variability, as well as a human error related to the failure to identify deviations.

A common target in medical imaging is trying to increase the speed of the diagnosis. Giving a correct diagnosis in the short time often helps patients in recovery better. Hence, solving these artefacts is fundamental for avoiding errors as well as increasing the speed. Moreover, in order to improve the speed of the diagnosis, we can resort to the use of a web-platform. Using a web-platform can bring many advantages for the research in medical image field such as 1) managing the images and the associated metadata needed for developing, testing and validating novel algorithms for medical image analysis, 2) increase collaboration between academia, industry and healthcare providers, 3) collaboration between multiple clinical institutions (share data), etc. Further, nowadays with the evolution of the deep learning techniques, there are operating system and processing power limitations which prevent applications from running on every type of workstation. By developing web-based tools, it is possible for users to access the medical image processing functionalities wherever the internet is available. Digital images should be processed, saved and retrieved easily and quickly using the software. They must compromise their characteristics in terms of reading, writing, and representing different image formats, applying various automatic analysis methods on the images in 2D and 3D, and applying the latest image processing methods to accurately segment and visualise the data (Chabat et al., 2000; Osteaux et al., 1992). These functionalities are necessary for computer assisted diagnosis and therapy.

In order to follow this line, we propose an approach that supports data consolidation and integration based on the well known XNAT[1] web-platform. XNAT is an open source imaging informatics platform, developed by the Neuroinformatics research group at Washington University (Herrick et al., 2016). It facilitates common management, productivity and quality assurance tasks for imaging and associated data.

1.1. Deep learning approaches in medical imaging

As 3D and 4D imaging are becoming routine, and with physiological and functional imaging increasing, medical imaging data is increasing in size and complexity. Therefore, it is essential to develop tools that can assist in extracting information from these datasets. Nowadays, the researchers are using mainly machine learning techniques to develop tools. Machine learning is a set of algorithmic techniques that allow computer systems to make data-driven predictions from large data. These techniques have a variety of applications that can be tailored to the medical field (Akkus et al., 2017). There has been a significant effort in developing classical machine learning algorithms for different

[1] https://www.xnat.org/

problems, for example segmentation of normal (e.g., white matter and gray matter) and abnormal brain tissues (e.g., brain tumors) in MRI. However, creation of the imaging features that allows to solve these problems requires careful engineering and specific expertise. Furthermore, traditional machine learning algorithms may do not generalise well. Despite a significant effort from the medical imaging research community, there are still unsolved problems due to many facts such as normal anatomical variations in brain morphology, variations in acquisition settings and MRI scanners, image acquisition imperfections, and variations in the appearance of pathology. But, an emerging machine learning technique referred to as deep learning (LeCun et al., 2015), can help avoid limitations of classical machine learning algorithms, and its self-learning features may enable identification of new useful imaging features for quantitative analysis of brain MRI. Deep learning techniques are gaining popularity in many areas of medical image analysis (Vasilakos et al., 2016), such as computer-aided diagnosis of brain disease, computer-aided detection of breast lesions (Kooi et al., 2017), computer-aided diagnosis of breast lesions and pulmonary nodules (Cheng et al., 2016), and in histopathological diagnosis (Litjens et al., 2016).

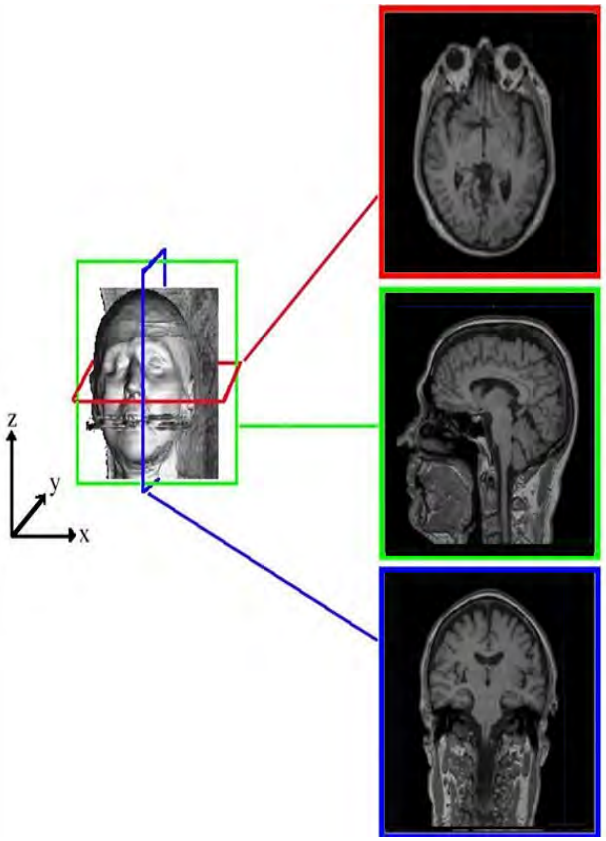


Figure 2: Axial, sagittal and coronal views are depicted in red, green and blue, respectively.

1.2. Objective

The main goal of this thesis is to develop a fully automated quality assessment pipeline for unsupervised brain MRI processing. After the preprocessing step, we treat several problems related to quality and orientations of the images. Furthermore, we integrate the fully automated pipeline into the XNAT platform. In views of the above information, realising tools for quality control of the images such as the integration of these in a web platform, it has become strictly necessary. For these reason, the workflow of this master thesis is organised as follow:

1. As the first step we propose an application for brain MRI images quality control, mainly based on deep learning: a Python script to organise, QC, and collaborate on neuroimaging processing.
2. Then we build a container with the idea of integrating this work into the installed web platform.
3. The last part is dedicated to the installation of the well-known web platform XNAT and the integration of the container into XNAT.

Moreover, we divided this master thesis into the several sub goals listed below:

Transfer learning approach. For reaching most of the prefixed goals we decide to base our work on a deep learning approach; in particular, we decide to use a transfer learning (fine tuning) technique. Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task. We use the VGG16 (Simonyan and Zisserman, 2014a) model pretrained on ImageNet in all the experiments. ImageNet is a research project to develop a large database of images with annotations, e.g. images and their descriptions (Krizhevsky et al., 2012). We decide to use transfer learning technique because usually training CNN's (Convolutional Neural Networks) from scratch is very complicated and time-consuming and it require always a significant amount of labelled data that are not always easy to retrieve.

Data. Another important objective prefixed in this thesis is to develop a method not related to the domain of the data (Ben-David et al., 2010). Nowadays, there are many problems related to the domain adaptation such as the use of different vendors MRI scanner or the different range of intensity for representing the same tissue. It is well known that CNN's methods for classification perform badly when training and testing data are drawn from different distributions. Regarding to this, our idea consists in to build a robust method that works in different domains. For doing that we build an algo-

rithm that works independently from MRI vendors or intensity range of the images. In order to achieve this goal, we use several different public datasets and, we always test our methods with cases from different scanners and image protocols during the training phase.

Ground truth generation. Working with CNN's needs a significant amount of labelled data. Even though we have plentiful of labelled training datasets, still there are many task where the ground-truth is not present. This is a common problem in quality control algorithms. For example, there is no labelled data in the literature for the classification quality of the brain skull stripping. The consequence of not having a sufficient amount of labelled data can lead to a failure of the CNN's in prediction time. One tendency in order to remedy this problem is to generate synthetic data starting from the real one, different being the strategies adopted by the researchers for this purpose. Consequently, we dedicate a part of this thesis to build methods for generate synthetic ground-truth in order to provide data needed by the different methodologies developed.

Experimental. In this thesis, we treat several problems related to the image quality control. When we load the medical images in a software is always ambiguous the storage of them in the memory. Normally, the information regarding the data storage are contained in the DICOM file (Digital Imaging and Communications in Medicine). DICOM is a specification for the creation, transmission, and storage of digital medical image and reports data. However, due to the different process applied to the medical images, the information contained in the DICOM can be corrupted or lost. For this reason, we build a method based on a deep learning approach that is capable to recognise the relation between the different views (axial, sagittal, coronal) and the axis of the matrix that will contain the data (see Figure 2). The second experiment that we conduct has the goal to identify the presence of the skull in the brain MRI images. Recognising the presence of the skull in the brain images in an automatic way can be helpful and time-saving when the number of images to process is big. However, more important then recognise the presence of the skull is knowing the quality of the skull-stripped image. For this reason, we dedicate the last part of this thesis to build a method which give us a quality index of the skull-stripped image.

Platform integration. As the last goal, we want to integrate the code into a web platform based on the XNAT software. For reaching this goal, we put the built script inside a container. After this, we integrate and plug this container into the XNAT platform.

2. State of the art

Most of the time, brain MRI image data in their raw form, are not immediately usable for applying deep learning algorithms and extracting biologically-meaningful information. Frequently images need to be preprocessed using a wide variety of software tools, optimised and combined together into a preprocessing pipeline. There are many excellent examples of preprocessing pipelines in the literature, such as those used by Glasser et al. (2013) or by Strother (2006).
Often pipelines are built using tools contained in FSL[2]. FSL is a comprehensive library of analysis tools for FMRI, MRI and DTI brain imaging data. FSL platform is a collection of different tools for medical images. Most of them are related to the preprocessing of the images such as FSL-BET (Smith, 2002), that has the aim of performing the brain extraction, or FSL-FLIRT (Jenkinson and Smith, 2001) used for the registration process. However, after the execution of the preprocessing step, there are not many tools to assess the success of the preprocessing algorithm and therefore the quality of the images' results. Hence, identifying bad preprocessed datasets using the actual tools is not always straightforward, and it can be prohibitively time consuming for large datasets.
Preprocessing protocols have been developed to extract metrics that can be viewed as a cohort-level summary from which outliers are selected for manual quality-assurance. Gardner et al. (1995) showed that human observers demonstrated poor sensitivity when evaluating intentionally degraded MRI volumes, as opposed to an automated approach, which detected even minimal noise in the images. One method to reduce the workload of visually inspecting a large number of MRIs is parallel processing by multiple investigators, such that each investigator examines a subset of the data. However, such an approach can be unreliable, as each investigator uses a different threshold for accepting or excluding data. Additionally, this approach is time-consuming, which makes the task of maintaining and updating the image quality information of large growing 3D-MRI datasets in a timely manner challenging.
Although it is possible to find different papers about each one of the problems mentioned before, there are very few studies in the literature that directly explore the general issues of the QC. For example, Bennett and Miller (2010) shows in his studies that poorly executed application (or lack thereof) can can compromise the trustworthiness of a study. The mentioned problems

[2] https://fsl.fmrib.ox.ac.uk/fsldownloads_registration

Table 1: Datasets used in this master thesis. The upper table is related to the Campinas- Calgary dataset, this dataset has been used only in the training phase. The table below is related with all the datsets used in the testing phase.

TRAINING DATASET					
Name	**Vendor**	**Field**	**Modality**	**Brain Mask**	**N. cases used**
CAMPINAS-CALGARY	Siemens	1.5 T	T1	yes	60
		3.0 T	T1	yes	60
	Philips	1.5 T	T1	yes	59
		3.0 T	T1	yes	60
	GE	1.5 T	T1	yes	60
		3.0 T	T1	yes	60
All					359

TESTING DATASETS					
Name	**Vendors**	**Fields**	**Modality**	**Brain Mask**	**N. cases used**
IXI-dataset	Philips,GE	3T/1.5T	T1/T2	no	581
WMH T1	Philips, Siemens, GE	3T/1.5T	T1/T2	no	120
OASIS2	Not specified	3T/1.5T	T1/T2	yes	365
ADNI2	Not specified	3T/1.5T	T1/Flair	yes	194
ATLAS R1.1	Philips, Siemens, GE	1.5T	T1	no	219
IBSR	Siemens, GE	1.5T	T1	yes	18
MICCAI 2016	Siemens, GE	1.5T	T1	yes	15
ISBI 2015	Siemens, GE	3.0T	T1/T2/Flair	yes	5
All					1517

can be compensate by the QC, hence QC it has became a critical issue to solve in brain imaging. This topic has been explored in the literature, although very often research is mostly focused on quality assurance rather than QC. Quality assurance is focused on avoiding the occurrence of problems by improving a process while QC is focused on finding possible problems in the output of that process. For example, the Function Biomedical Informatics Research Network (FBIRN) set of recommendations (Glover et al., 2012) is solely focused on quality assurance. In a similar vein, Friedman and Glover (2006) explore an interesting set of quality metrics, but they focus on stability, signal-to-noise ratio (SNR), drift, and other hardware performance issues related to MR scanners, not specifically on the type of artefacts that can be found in MR imaging even when complying with the best quality assurance policies.

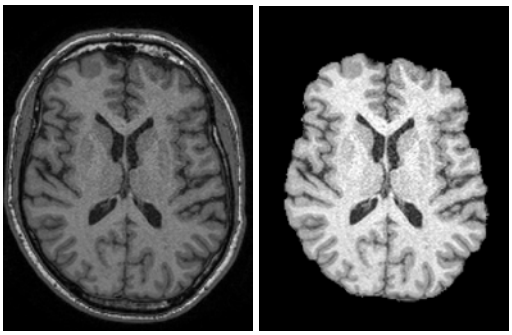
3. Materials and methods

3.1. Datasets

For addressing the different problems treated in this thesis we decide to use MRI (magnetic resonance image) from nine public datasets (see Table 1). We make this decision because we have the necessity of building methods that are robust to the change between domains. In order to do this, we decide to use a separate set of data for the different phase of training and testing. It has to be noticed that in training phase we only use the Dataset Calgary-Campinas (Souza et al., 2017), but we lately test with all the others datasets to prove the generalisation of the methods. All the datasets used are listed below:

1. **Campinas-Calgary 359 dataset**
   The dataset Campinas-Calgary 359[3] is composed of images of healthy adults (29-80 years) acquired on scanners from three vendors (Siemens, Philips and General Electric) at both 1.5 T and 3 T. CC-359 is comprised of 359 images, 60 subjects per vendor and magnetic field strength. The dataset is approximately age and gender balanced, subject to the constraints of the available images. It provides consensus brain extraction masks for all volumes generated using supervised classification. Manual segmentation results for twelve randomly selected subjects performed by an expert are also provided. The CC-359 dataset allows investigation of 1) the influences of both vendor and magnetic field strength on quantitative analysis of brain MRI; 2) parameter optimisation for automatic segmentation methods; and potentially 3) machine learning classifiers with big data, specifically those based on deep learning methods, as these approaches require a large amount of data (Souza et al., 2017). Figure 3 shows one sample

[3] http://miclab.fee.unicamp.br/calgary-campinas-359



(a) (b)

Figure 3: Example of the same slice before and after applying the mask of the brain-extraction. The slices come from the dataset CC359. a)Slice with the skull. b)Slice skull-stripped.

of the CC359 before applying the brain extraction mask and after applied it.

2. **IXI dataset**
   The Information eXtraction from Images (IXI)[4] dataset, is a collection of structural MRIs from 581 healthy adults across the lifespan (20−86 years old). The IXI dataset was collected in 2005/2006 from three sites in the UK (each with a different scanner system) and includes T1, T2, Proton Density(PD), and MRA images. Here we only used the T1 structural images. The dataset is freely available from the mentioned website. The IXI dataset has been used in numerous studies investigating structural properties of the brain and related differences due to healthy aging (Ardekani and Bachman, 2009; Madan and Kensinger, 2016; Zhang et al., 2014).
3. **White matter intensity dataset MICCAI 2017**
   Image data used in this dataset are coming from Medical Image Computing Computer Assisted Intervetion (MICCAI)[5] challenge 2017. The images were acquired from five different scanners from three different vendors in three different hospitals in the Netherlands and Singapore. For each subject, a 3D T1 image and a 2D multi-slice Fluid Attenuated Inversion Recovery (FLAIR) image are provided. The manual reference standard is defined on the FLAIR image. From this dataset we have used 120 cases.
4. **OASIS2 Dataset**
   The Open Access Series of Imaging Studies (OASIS2)[6] is a series of magnetic resonance imaging data sets that is publicly available for study and analysis. This dataset consists of a longitudinal collection of 150 subjects aged 60 to 96. Each subject was scanned on two or more visits, separated by at least one year for a total of 373 imaging sessions. For each subject, 3 or 4 individual T1 MRI scans obtained in single scan sessions are included. The subjects are all right-handed and include both men and women. 72 of the subjects were characterised as nondemented throughout the study. 64 of the included subjects were characterised as demented at the time of their initial visits and remained so for subsequent scans, including 51 individuals with mild to moderate Alzheimers disease. Another 14 subjects were characterised as nondemented at the time of their initial visit and were subsequently characterised as demented at a later visit. In the description of this dataset, there are no informations about the vendor machine used. From this dataset we have used 365 cases.
5. **ADNI-2 dataset**
   In the Alzheimer's Disease Neuroimaging Initiative (ADNI-2)[7] as explained by Jack et al. (2008) researchers collect several types of data from study volunteers throughout their participation in the study. Data collection is performed using a standard set of protocols and procedures to eliminate inconsistencies. Subjects in the baseline ADNI-2 dataset have 1.5T and 3T T1 structural MRI data. The dataset included a total of n = 5,738 scans acquired 3, 6, 12, and 24 months from the following participants: 198 healthy controls, 111 individuals with significant memory complaint, 182 individuals with early mild cognitive impairment, 177 individuals with late mild cognitive impairment and 155 probable Alzheimer's Disease patients. For our experiments, we pick a subset of 194 patients all coming from the healthy control class. Due to its vastness, no informations about vendors machines are furnished
6. **ATLAS R1.1**
   The Anatomical Tracings of Lesions After Stroke (ATLAS R1.1)[8] dataset, is an open-source dataset of 304 T1 MRIs (Liew et al., 2018). 304 MRI images from 11 cohorts worldwide were collected from research groups in the ENIGMA Stroke Recovery Working Group consortium[9]. Images consisted of T1 anatomical MRIs of individuals after stroke. These images were collected primarily for research purposes and are not representative of the overall general stroke population. For each MRI, brain lesions were identified and masks were manually drawn on each individual brain in native space using MRIcron[10].
7. **IBSR dataset**
   Internet Brain Segmentation Repository

[4] http://brain-development.org/ixi-dataset/
[5] https://sites.google.com/site/braintumorsegmentation/
[6] https://www.oasis-brains.org/data
[7] http://adni.loni.usc.edu/adni-go-adni-2-clinical-data-available/
[8] http://fcon_1000.projects.nitrc.org/indi/retro/atlas.html
[9] http://enigma.ini.usc.edu/ongoing/enigma-stroke-recovery/
[10] https://www.nitrc.org/projects/mricron

(IBSR18)[11] dataset which is one of the standard datasets for tissue quantification and segmentation evaluation. The dataset consists of 18 MRI volumes including: ten volumes for training , five for validation and three for testing. For the training and validation images, the corresponding ground truth (GT) is provided, while for the testing set it will not be available. The number of valaible images as suggetested by the name are 18.

8. **ISBI 2015**
   International Symposium on Biomedical Imaging (ISBI)[12] dataset consist in a longitudinal studies of multiple sclerosis (MS) patient. The images were given by the 2015 Longitudinal MS Lesion Segmentation Challenge. The image are acquired in T1 modality with a magnetic field of 3 tesla. Due to the bad quality it was possible use just seven cases of this datasets.
9. **MICCAI 2016**
   The MICCAI 2016[13] is composed of 15 training scans acquired in different image domains: 5 scans (Philips Ingenia 3T), 5 scans (Siemens Aera 1.5T) and 5 scans (Siemens Verio 3T). For each subject, 3D T1 MPRAGE, 3D FLAIR, 3D T1 gadolinium enhanced and 2D T2/DP images were provided, presenting different image resolutions for each image domain (see the organiser's website for the exact details of the acquisition parameter and image resolutions). Manual lesion annotations for each training subject were provided as a consensus mask among 7 different human raters.

3.2. Methods

For the problem treated in this master thesis, a transfer learning approach has been conducted by using an already pretrained and well-know VGG16 architecture (see Figure 4). Transfer Learning freezes the bottom layers of the CNN's to extract image features vectors from a training set in a different domain, which can then be used to train a new classifier for this domain. The strategy here involves the use of a pre-trained VGG16 network, developed by Simonyan and Zisserman (2014b), as an image feature extraction technique. The weights that we use are the ones extracted from the VGG16 trained on ImageNet (Deng et al., 2009). Even if the medical images may appear to be very different from the images used in ImageNet, Tajbakhsh et al. (2016) recently showed the potential for knowledge transfer to the medical imaging domain.
Transfer Learning begins with copying (transferring) the weights from a pre-trained network to the network we wish to train. The exception is the fully connected

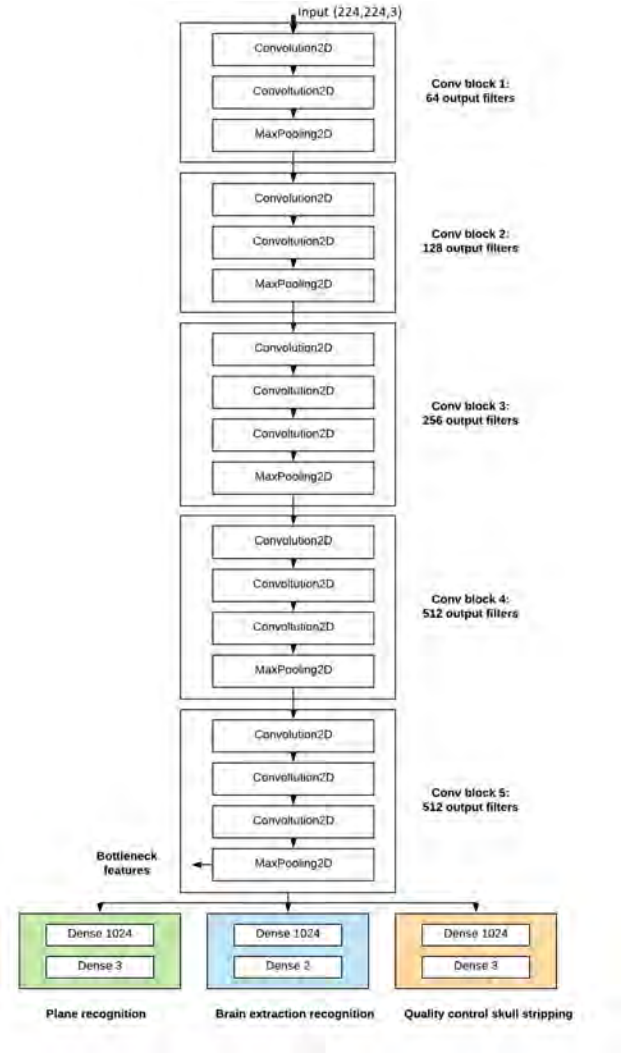


Figure 4: VGG16 archtectures. The block in different colors indicate the fine tuning of the differents modules.

layers, especially the last one whose number of nodes depends on the number of classes in the dataset. A common practice is to replace the last fully connected layer of the pre-trained CNN with a new fully-connected layer that has the many neurons as the number of the classes in the new target application. In our study, we deal with two and three class classification tasks. Therefore, the new fully connected layer has two or three neurons depending on the application under study. After the weights of the last fully connected layers are initialised, these layers will be fine-tuned with the training dataset. In general, the early layers of a CNN learn low level image features, which are applicable to most vision tasks, but the late layers learn high-level features, which are specific to the application at hand. Therefore, fine-tuning the last few layers is usually sufficient for adapt the CNN to a new task. Hence, the fully connected layers of our pretrained architecture (VGG16) were replaced with a new fully connected layer, and the

[11] https://www.nitrc.org/projects/ibsr/
[12] http://biomedicalimaging.org/2015/
[13] https://portal.fli-iam.irisa.fr/msseg-challenge/overview

labelled data were used to train only the added layers while keeping the rest of the network the same.

3.3. CNN's settings

The pre-trained VGG16 on Imagenet (Simonyan and Zisserman, 2014a) consists of approximately 15 milion of parameters trained using 1.2 milion images labeled with 1000 semantic class. In the strategy adopted, we decided to freeze the weights of all the convolutional layers, that were around 14 milions of parameters, and we trained the fully connected layers. We decided to add two fully connected layers one with a fixed size of 1024 and the other with a size depending on the treated problems. Thus, the trainable parameters were around 500000.

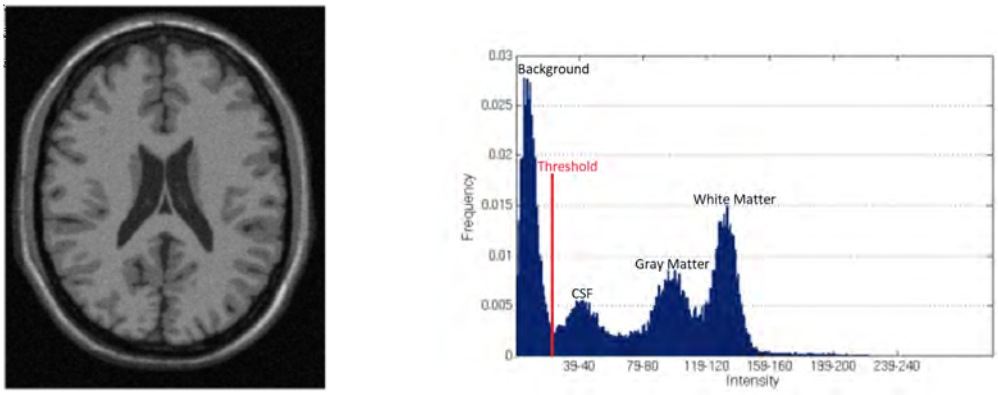
For the training phase we decided to use an early stopping technique approach, hence the number of training epochs was stopped when a condition was verified. In machine learning, early stopping is a form of regularisation used to avoid overfitting when training a learner with an iterative method, such as gradient descent. Early stopping rules provide guidance such as how many iterations can be run before the learner begins to over-fit. In the method proposed, we decided to compute a validation step every epoch. If there was no improving in the accuracy performance over six sequential validations, the training of the net was stopped and the weights of the best validation accuracy epoch stored. We set the batch size to 32. It has been observed in practice that when using a larger batch there is a significant degradation in the quality of the model, as measured by its ability to generalise (Ren et al., 2015). Moreover, the lack of generalisation ability is due to the fact that large-batch methods tend to converge to sharp minimisers of the training function (Keskar et al., 2016). Furthermore we decided to use Adam optimiser (Kingma and Ba, 2014). In choosing an optimiser what is important to consider is the network depth (benefit from per-weight learning rates if the network is deep) and the type of layers and the type of data. For deciding the learning rate of the trainable layers we conducted an investigative analysis. The parameter that we choose ensured convergence for all the applications. Hence, following on from what we said, we choose as a final parameters a learning rate equal to 10^{-4}, batch size of 32, max number of epochs to 50 a patience epochs equal to 6 (early stopping condition). As images input, we used the MRI slices without doing any division in patches, but we re-sliced the input images according to the input layer of the VGG16 (224x224x3). Since the VGG16 architecture receives color images as its input in two of the problems, we simply repeated the first channel and produced 3-channel RGB-like images, and in the remain problem we decided to adopt another strategy for generate the 3-channels images. In the next sections we will analyse in detail the individual modules of the framework.

3.4. Plane orientation recognition

The planes of MRI brain images are categorised into three different groups axial, sagittal and coronal. For understanding the relationship between the different acquisition plane and the axis of the variable where the images are stored we usually rely on the information contained in the Neuroimaging Informatics Technology Initiative (Nifti) file header. However, the header's information sometimes may be corrupted due to different reasons such as anonymisation process or use of different protocols. This may cause possible errors in future applied algorithms. In this module, we built an algorithm able to individualise the different views of Brain MRI images. An important point to stress is that in these methods we only use images features, similar to humans, so the network is somehow learning to identify the brain shape to recognise the plane. The strength of this method is that we do not need parameters to tune the learning process.

For all the experiments, we used the Campinas-Calgary dataset in training phase and all of the others datasets for the testing phase. For preparing the training data, according to the description of the dataset, all the images were acquired within the same protocol. In the protocol of medical images, it is possible to find all the information about the orientation, number of slices, views etc. Furthermore, an extra visual check was performed for avoiding errors. As a consequence of the above, we can assume to find the same views on the same axis for all the training cases. After this step, we built the training dataset by extracting slices from the volume along its three different planes.

In the initial experiment, we trained the CNN with all the slices of each case. However, due to the similarity between the extremes slices in the different views, the results obtained were not good in terms of accuracy and time-consuming. So, due to the anatomy of the brain, where is clear that in the centre of each single views is possible to find more structure, we can affirm that the most informative slices for this problem are the ones localised in the centre of the MRI images series. Hence, it is unnecessary in terms of time to use all the slices per view. Nevertheless, define where a central slice is in a volume with no prior information is not an easy task. For this reason, we build a function able to extract the central slice index of all the plane (see Algortihm 1). This function is based on counting the number of nonzero pixels over all the slices and assume as a central slice the one with the maximum number of it. However, in the first approach tried we noticed that the black area of the images was not set to zero as expected. Commonly this problem is caused by the noise introduced during the acquisition time. Because of this the procedure described above was failing. So we decided to add an extra step, and before counting the nonzero pixels, we performed a histogram thresholding. Assuming that the background occupies a large



(a) Axial slice of brain MRI.

(b) Histogram correspondent to the MRI slice.

Figure 5: Histogram of the intensity distribution of a brain MRI images. The first mode value after the zero value represents the background of the image.

Algorithm 1 Central slice localisation

```
procedure
    for each slice S in the volume do
        Hist ← histogram of the slice
        Threshold ← value after the first mode
        for each pixel in S do
            if S(x, y) < Thresold then
                S(x,y) ← zero
            else
                S(x,y) ← one
    central idx slice ← slice max amount of ones
```

part of the slices and that the mean value of the noise distribution has values close to the zero value, thresholding the histogram in the first minimum after the first big mode, will eliminate the noise (See Figure 5). The MRI histogram was computed as follows: given the intensity value of the native space MRI, the range of the intensities $[0, I_{max}]$ was divided into 30 bins of width $I_{max}/30$, where $I_{max} = max(I(x, y, z))$.
In order to allow the CNN to work in a more general way, we decided to use images coming from both skull-stripped and not skull-stripped cases. The skull-stripped images were obtained through the multiplication of the given mask and the native volume of the Dataset CC-359. Then, as a last step before starting the training phase, we performed the normalisation of all the volumes. Successively, for building the training set, we extracted a range of 10 slices around the centre slice of each plane. Hence, as a training test, we collected 12000 slices, equally distributed in the three class axial, sagittal and coronal. We decided to split the dataset into two parts composed of 80% (training) and 20% (validation). This division was conducted to assess the conditions of the early stop technique.
For the evaluation part, we used all the others datasets illustrated in Table 1. For a correctly use of these datasets, we individualise the location of the different views by the use of the header, and we ensure the correctness of this information by visual inspection of each one.

3.5. Brain-extraction recognition

Our motivation for devising algorithms for recognising if the MRI is skull-stripped was multifaceted. We aimed at establishing a method that requires no parameter tuning and handles images coming from different clinical routines. Furthermore, we aimed at building solid algorithms capable to work with all the modalities acquired (T1, T2, PD, FLAIR).
Even for a human recognising the presence of the skull is an easy task, but when the number of cases increases analysing all of it can become a waste of time. For this reason, often the researchers do not perform the verification in all the cases, but they analyse just a few samples picked from the whole datasets. However, it can happen that some cases are not uniform with the rest of the dataset. Moreover, usually the data analysed in the different medical image algorithm do not come from the same datasets. Accordingly with the different reason explained, it can happen to misclassified some cases as skull-stripped when they are not, or the opposite. This misclassification can lead to a failure of the following steps applied to the dataset, with the risk of doing an erroneous diagnosis.
As a first step for these methods, we applied the mask to the 359 cases, and we took the result has a ground truth for the skull-stripped class. Thus, we built the two-class data needed for the training of this problem as follow:

- Skull-stripped Brain MRI (359 cases)
- Not skull-stripped Brain MRI (359 cases)

Even if is easily possible extend this problem to all the views (axial, sagittal, coronal), we decided to deal with this problem in a slice oriented way; for this reason we extracted only the axial views slices to train

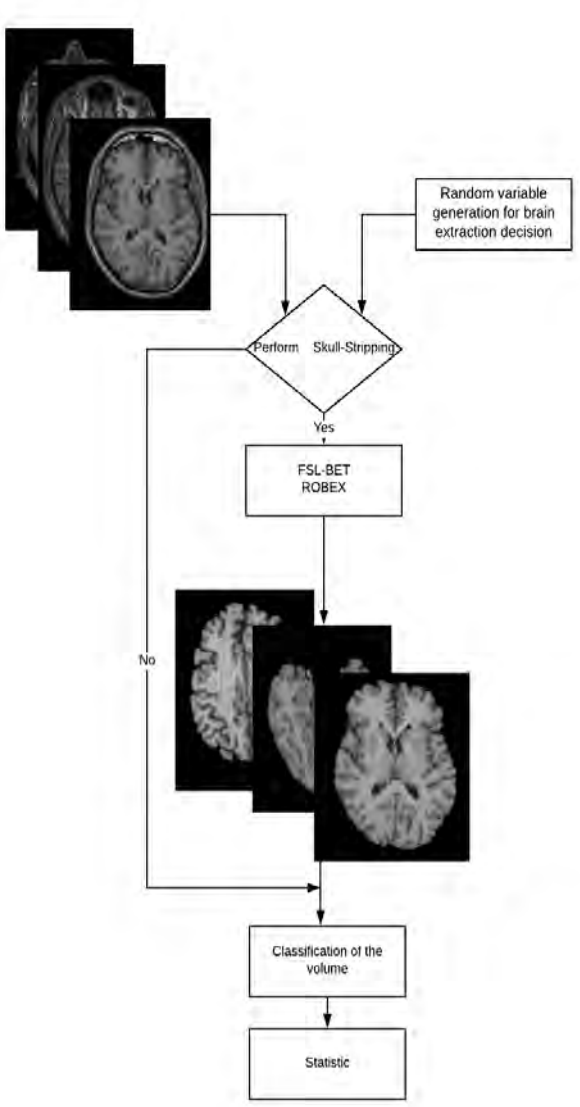


Figure 6: Procedure used for perform the testing phase of the brain-extraction recognition.

the net with. For checking the axis where the axial was stored, we used the first module developed (plane recognition). This module is described in the section 3.4. Before extracting the slice, we perform the mean shift and division by the standard deviation by volume for homogenise the data (Li, 2012). The idea is to allow different data sets to be comparable. Following, we splitted these data as follow: 1) 80% for the training phase, 2) 20% for the validation phase. Lately, we tested the CNN in all the others datasets. For performing the testing phase, we built a function that receive in input only cases with the skull. Successively, we generated a random binary variable that indicated whether or not to apply the brain extraction algorithms. As a next step, we applied the net for predict the class of the volumes, and we compare it with the expected classification (see Figure 6).

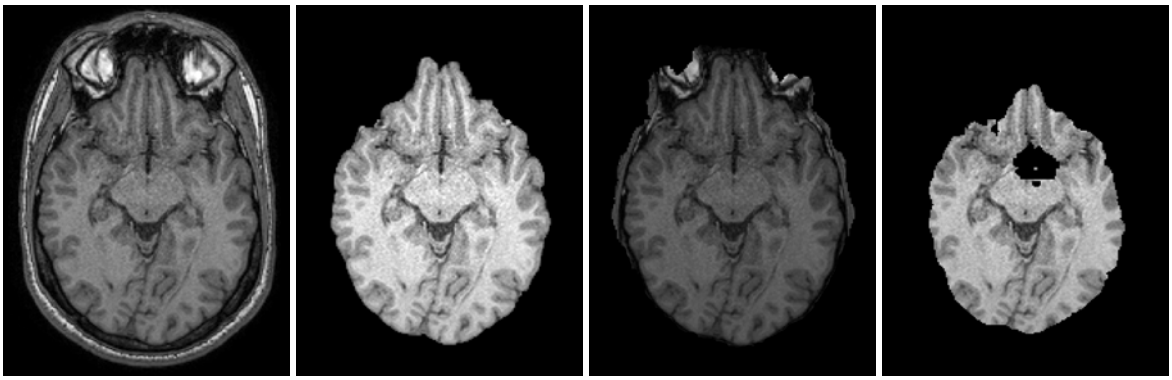
3.6. Quality control of the skull stripping

When a brain extraction is performed analysing the quality in an automatic and unsupervised way, it can be an interesting contribution to the researcher's community. Nowadays there are different public tools used for accomplishing the brain extraction, for example, the most used are FSL-BET (Jenkinson et al., 2012), Robex (Iglesias et al., 2011), BEaST (Bouckaert et al., 2014) etc. In some of these tools, there is the possibility to tune some parameters to better adapt the tool to each particular case. But, tuning these parameters require a large amount of time. Thus, most of the time these tools are used with the standard configuration, with the consequence of not having guaranteed in the quality of the result.

Hence, the primary goal of this module is to give a measure of the quality of a skull-stripped image. Besides, we can use the index given in output by the CNN to build many different tools; for example, one idea is to integrate this module in a pipeline for building an automatic and unsupervised method to tune the parameters of the skull stripping.

From the original Calgary Campinas dataset we used the given mask of the brain extraction as a ground-truth. Lately, for building the bad skull stripping cases, we decided to make an automated script for corrupting the correct data mask. In this script, we used morphological operations such as dilation and erosion to corrupt the original masks (ground-truth). The corruption is performed in a non uniform and controlled way. The purpose of this decision was to corrupt the data more realistically. In doing so, we can control what are the parts to include in the bad skull-stripped case, for example, eyes, part of the skull neck and other parts. In order to follow this line, we built different morphological structuring elements, and lately, we used them during the dilation and erosion operations. It is important to highlight that, an essential part of the morphological operation such as dilation and erosion is the structuring element used to probe the input image. A structuring element is a matrix that identifies the pixel in the image being processed and defines the neighborhood used in the processing of each pixel. Typically a structuring element is chosen according to the deformation that we want to have in the input image. In addition, for generating more realistic bad skull stripping cases, we decided to generate skull stripping volumes starting from the original cases by applying a brain extraction tool (FSL-BET) herein the parameters were tuned wrongly. All of these cases were visually checked for ensuring the belonging class. One example of the distorted images are shown in Figure 7 After that we have conducted two main experiments that are described below:

QC in the original space: In the first experiment, we decided to train the CNN with the image in their native space. As the first step of this experiment, we



(a) (b) (c) (d)

Figure 7: Example of distortions applied on the data. a) Original Slice. b) Slice good skull-stripped. c) Slice corrupted by dilation of the mask. d) Slice corrupted by erosion of the mask.

performed the mean shift and division by the standard deviation. After that, we proceed to built the training set by extracting all the non-black slice from the generated dataset. Before proceeding with the training, we checked if all the slices were extracted in the axial views. So in total, we extracted 70000 slices. As in all the previous experiments, we replicated the same slice in the three channels of the VGG input.
For evaluating the CNN, we used the datasets exposed in the section 3.1. From all the datasets we selected just the one in which a brain mask extraction was given. So after considering these masks as good skull-stripped, we performed the prediction on it using the CNN, and we finally checked the correctness of the prediction. Moreover, for testing the CNN in recognising the negative class, we corrupted the ground-truth in the same way performed in training. As a result, the CNN was able to recognise the bad skull-stripped case in which were present more morphological tissue than the expected one (dilated mask) for example, eyes part of skull etc. But it was not the same for the cases were the bad skull-stripped were caused by the presence of less tissue of the brain (eroded mask). For this reason, we decided to implement another method.

QC in the MNI space:

In this experiment, we decided to register all the cases of Campinas-Calgary dataset to the Montreal Neurological Institute (MNI) 1mm space. The MNI defined a new standard brain by using a large series of MRI scans on normal controls. To perform the registration we used FMRIBs Linear Image Registration Tool (FLIRT) (Jenkinson et al., 2002). The alignment to the MNI standard space was completed using the affine transformation process provided by FLIRT, which has 12 parameters or degrees of freedom. The affine process was employed to keep the shape and proportions of each brain. The 12 parameters are divided into the following subgroup: translation, shear, rotation and scale. Moreover,

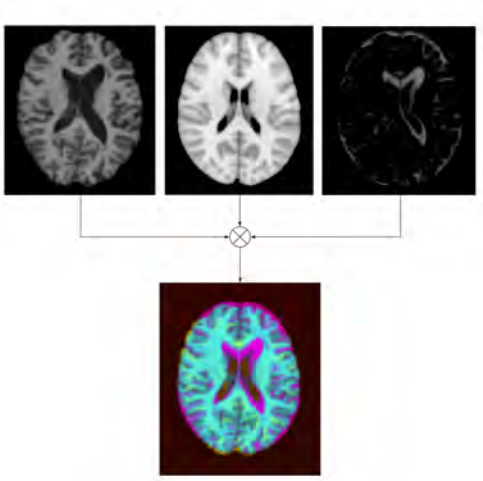


Figure 8: Input image of the VGG

the FLIRT algorithm gives the possibility to set others parameters such as interpolation or the reference volume. In our case, we chose a trilinear interpolation and the MNI152 template as a reference volume.
The decision to move into the MNI space was guided by the idea to put in the second channel of the VGG input image the correspondent slice of the mean volume MNI and in the third channel the difference between the slice in analysis and the MNI slice. In the Figure 8 is shown one example of the built image. We took this choice for allowing the CNN to capture the difference between the MNI volume and the slice analysed; this difference can help in detecting the skull stripping quality. In the previous approach, the CNN was not able to detect errors caused by the excessive erosion of the brain structure. The motivation behind this error was mainly due to the similarity between the upper slice of the brain and the extra eroded one. Hence, the CNN was getting lost in differentiating these slices. We solve this error by using the mean volume of the MNI as a metric of the expected volume size. We put in the third channel the difference

between the two volumes in order to reinforce this difference. Thus, with the new approach, we solved the issue of failing when the skull stripping tool was eroding too much the brain structures.
For testing this methodology, we used all the public datasets containing a manual brain mask and the ones in which was given only the segmentation mask of the different classes, where we reconstructed the brain mask by merging all the classes labels. In this approach, the final decision is based on the overall scores obtained by computing the mean over all the slices. However, it has to be noted that for each slice the CNN gave as output the probability of belonging to one of the two classes. Thus, it can happen that the prediction of the volumes can result in a good skull stripping class even if it includes some bad skull-stripped slices. For dealing with this problem different are the strategies that can be adopted such as correcting the bad slice and make the prediction again, or we can put a threshold of assessment under which the result of the skull stripping is not acceptable. After that, as an example of application, we built a function for automatically tune the parameters of FSL-BET for having the best skull stripping result. The scheme of this function is explained in the subsection 4.4. Lately, we tested this function on the WMH dataset and, we visually checked the correspondence between the results of the skull stripping, in which the parameters were varied, and the quality index produced by the CNN.

3.7. Implementation

The architecture that we use is a Tensorflow implementation of VGG 16. TensorFlow, developed by Abadi et al. (2015), is an interface for expressing machine learning algorithms and an implementation for executing such algorithms. Furthermore is an open source software library. For the fine-tuning scenarios, we used the pre-trained VGG16 model provided by Keras (Chollet et al., 2015). Keras is a high-level neural networks API, written in Python and capable of running on top of TensorFlow.

4. Results & Discussion

We now provide a qualitative and quantitative analysis of the performance of the developed methods. As an additional experiment, we conducted a test for evaluating the performance of the implemented pipeline for autotune the skull stripping tool parameter. For each of the methodology exposed in the section before we used various combinations of the public datasets. The cases selected for the different experiments were taken according to the information given by the description of the public datasets such as the presence of brain mask or number of axial slice.

Table 2: Plane Recognition Test. The results show in the table are computed by classifying the content of the three different axis of the volume. If one of these views is not recognised correctly all the volume is classified as wrong.

PLANE RECOGNITION TEST			
Name	**N. cases used**	**Acc. %**	**N. correct classified**
IXI	580	99.5%	578
WMH	120	100%	120
ADNI2	194	95.3%	185
ATLAS R1.1	219	100%	219
OASIS2	365	97%	355
General	1479	98.5%	1457

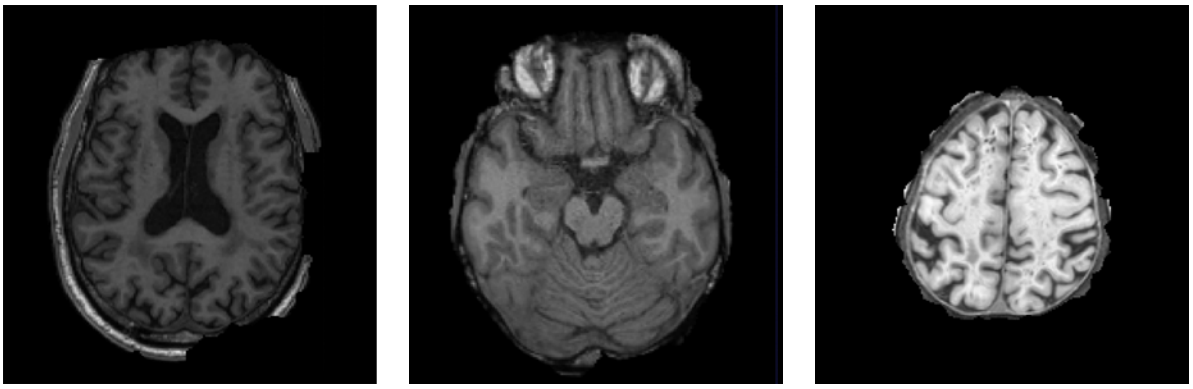
4.1. Plane recognition results

As a first step, we tested the methods plane recognition. This method has been tested on a total of 1479 cases coming from different datasets. More detail about the dataset are exposed in the Table 1. Due to the absence of ground truth, we first conducted a visual inspection procedure in order to generate labelled data needed for these experiments. After that, we computed the accuracy of the correctly predicted MRI cases when compared to the visually inspected category. The method reached a mean accuracy of 98.5% (n = 1457) of correctly classified cases. More detail about the experiments result are reported in Table 2.

Due to the strong difference between the three different views axial, sagittal and coronal this method achieved really good accuracy in prediction. However, in some case, the method developed was not able to correctly classify. This misclassification was mainly due to the failure on individualising the central slice in the volume. This error happened because the methods developed for extracting the central slice mainly relied on a histogram thresholding. Hence, if the noise of the background was relevant we were not able to isolate the structural part of the MRI volumes from the background. Thus, the central slice of the volumes was wrongly detected and as a consequences the CNN was getting confused in classify this no central slice. In the way that we performed the training the CNN was mainly memorising the shape of the central slice of the three different views, hence it was expected the misclassifications of slices different from the central. Moreover, in ordered to decrease the classification error we performed the prediction in all the three axis separately and in the end we did a majority voting between the three vector of probability obtained.

4.2. Brain extraction recognition

As a second stage, we evaluated the performance of the developed methods for recognising the presence of the skull in the MRI case. Because of the automatic methods developed for testing the CNN explained in sec. 3.5, it was possible to analyse all the datasets where the original images with skull were present. Hence, we



(a) (b) (c)

Figure 9: Example of cases miss-classified by the brain extraction method. It is clearly evident the presence of most of the skull.

Table 3: Brain extraction recognition test. The result show in the table are a obtain by computing the mean over all the volume inside each single dataset.

BRAIN EXTRACTION TEST			
Name	**N. cases used**	**Acc. %**	**N. skull-stripped (BET)**
IXI-dataset	581	100%	290
WMH T1	120	99%	60
OASIS2	365	100%	150
ADNI2	194	100%	95
ATLAS R1.1	219	100%	105
IBSR	18	98%	9
MICCAI2016	15	100%	7
ISBI	5	97%	3
General	1517	99.25%	719

Table 4: Testing results on the native spaces and in the MNI space. The accuracy is computed by classify each volume as true of false and then compare this with the ground-truth.

QUALITY CONTROL TEST (NATIVE vs MNI)			
Name	**N. cases used**	**Acc. native space**	**Acc. MNI space**
OASIS2	60	89%	95%
ADNI2	40	90%	93%
IBSR	18	100%	100%
MICCAI2016	15	100%	100%
ISBI	5	75%	70%
General	138	90.8%	91.6%

conducted an exhaustive test on over 1500 cases (see Table 3). As expected in this problem, we reached a really high accuracy over all the cases of 99.25%.
The presence of the skull in one image was easily recognisable by the CNN. This was due to different reasons such as the strong difference between the intensity of the skull and the intensity of the other structures in the brain or the mean volume size. Hence, many are the distinct features that allow the CNN to clearly distinguish the two classes. However, even if the problem was not considered very difficult some errors of classification were still present. These errors were due to the bad skull-stripping result generated by the tools (FSL-BET) that we used in the testing pipeline. To understand why these errors happened we decided to show some cases that were missclasified (See Figure 9). These images were examples of slices skull-stripped by the use of FSL-BET with the standard parameters. It was easy to see that the result of the skull-stripped slices was totally ambiguous. Due to the partial belonging to both of the classes, it was not an easy task classify these images in an automatic way. It was clearly visible that in the slice in Figure 9a was present more than half of the structure of the skull, even if the brain-extraction was already performed on it. Moreover, is important to highlight that these methods concern only to distinguish skull-stripped or not skull-stripped brain MRI. Thus, the case of bad skull-stripping are totally new for the CNN (not used in training phase) and so this leads to generate confusion in the prediction of them.

4.3. Quality control

For the evaluation of the performance of this last module, we used all the datasets where a brain mask was given. Also, we used the datasets in which the segmentation label of the different structure was present. From these, we generated a whole brain segmentation mask by merging all the label of the different class. We split the experiments into two different parts. First, we evaluated the performance of the CNN by treating the problem as a binary classification. Hence, the prediction was done by considering a threshold and the mean probability computed over all the slices for each case. As a result, we classified as true (good quality) the cases where the mean probability computed over all the slices was above the threshold and we set as false (bad quality) in the opposite case. The threshold can be set according to the required quality of the problem treated. In these experiments we set a threshold of 0.5. Moreover, we conducted the experiment for both of the methodologies developed: the one in the native space and the one in the MNI space. Results of these experiments are shown in Table 4.
Second, we decided to analyse deeper the behaviour of

Table 5: Result of the performance of the CNN tested on some cases. The probability is a mean computed over all the slice of each volume.

QUALITY CONTROL TEST		
Name dataset	**Case name**	**P(good skull-stripping)**
IBSR	07	0.95
	08	0.93
	09	0.97
	11	0.94
MICCAI	01016SACH	0.96
	01038PAGU	0.94
	08027SYBR	0.95
	08029IVDI	0.93
ISBI	01	0.60
	02	0.72
	03	0.30

our methodologies. Hence, we checked and reported individually some cases after registering them into the MNI spaces (see Table 5). From the result showed in Table 4 is possible to see that, our method for quality control as a binary classification, was able to achieve high accuracy. However, it may happen that one volume classified as a good quality contains some slices totally corrupted. For example, in Figure 10 is shown a case from the dataset MICCAI 2016 in which the overall probability to be classified as a good skull-stripped case was 0.86 but, when we analysed the probability of each slice was possible to realise that two of them were corrupted by the presence of one eye. For this reason, we decided to investigate more the probability of each slice individually rather than considering the overall mean probability. The main idea is to communicate to the user the index of the slice that can be corrupted even if the case has been classified as good. However, in Figure 10 is evident that the skull stripping was well performed (except for the first two slices) and, according with this, the overall accuracy of the prediction was really high. Furthermore, by analysing more in detail the result of some cases we realised that in the native space our algorithm was not able to recognise an extra erosion of the brain. This problem appears because a slice extra eroded can be easily confused as one located in the upper part of the brain. These slices according to the typical anatomy are smaller than the central. For avoiding this error, we decided to move in the MNI space in order to have a reference of the standard size of a brain.

Lastly, we computed an extra test by corrupting the ground-truth. In Table 6 we show the result of the test and is clearly understandable that in the MNI space the accuracy regarding the eroded ground-truth is much better than the one in the native space. This happens because this methodology, thanks to the data structure built as input (image in the first channel, MNI image in the second and difference in the third channel), was able to capture the difference between the mean volume MNI and the slice. Assuming that the there is no big variation on the volume between individuals we decided to use the mean volume MNI as a reference. In this way we created a metric for evaluating the corruption of the skull stripping. Hence, thanks to this metric of evaluation we reached good accuracy even when the cases were of bad quality due to the excessive erosion. However, some error may appear when a brain is totally different in terms of size from the mean standard MNI volume. For example, in many different diseases such as multiple sclerosis, Alzheimer or stroke it can be present the brain atrophy generation. Brain atrophy, or cerebral atrophy, is the loss of brain cells called neurons with the consequent reduction in brain volume. The atrophy is strictly related to the ventricles enlargement, hence when there is big atrophy the difference between the MNI volume (healthy) and the case in the analysis can be substantial. According to the method developed, the big difference between the two volumes is interpreted as bad quality of the skull stripping. But in this case, the difference between the two volume is not related to the quality of the skull stripping. For this reason it is considered a limitation of our methods. However, for avoiding this problem we can make a crop of the ventricles because usually the atrophy is mainly concentrated in the ventricles of the brain.

4.4. Application example

To show the potential and the efficacy of our framework we built an application that iteratively autotune the parameter f of the tool FSL-BET. None of the tools used for the preprocessing step, incorporate a built-in quality control. Thus, the only way to control the integrity of the volume is performing a visual inspection. Furthermore, most of the time these tools give the possibility to tune some parameters. The possibility to tune the parameters is given in order to generalise better these tools and let them work in a larger scenario of images. However, tune these parameters can be tedious and a dispendious of time. For this reason, we built an application that allows the autotuning of these parameters. The key point of our work is that the different applications developed can be combined together to the user's preferences. Hence we can create different pipelines. In Figure 11 it can be seen how different modules are used together for reaching one objective that in this case is the autotune of the parameters. This is just one of many possible pipelines that can be realised with the different modules of the framework developed. In this pipeline we first check in which axis the axial view is stored, then we analyse the slice for recognising if the MRI volume is already skull-stripped. In case is not skull stripping we proceed in applying one of the standard algorithms for brain extraction that we mentioned before. After this step, we register the image to the MNI space, and through the use of another CNN we give an

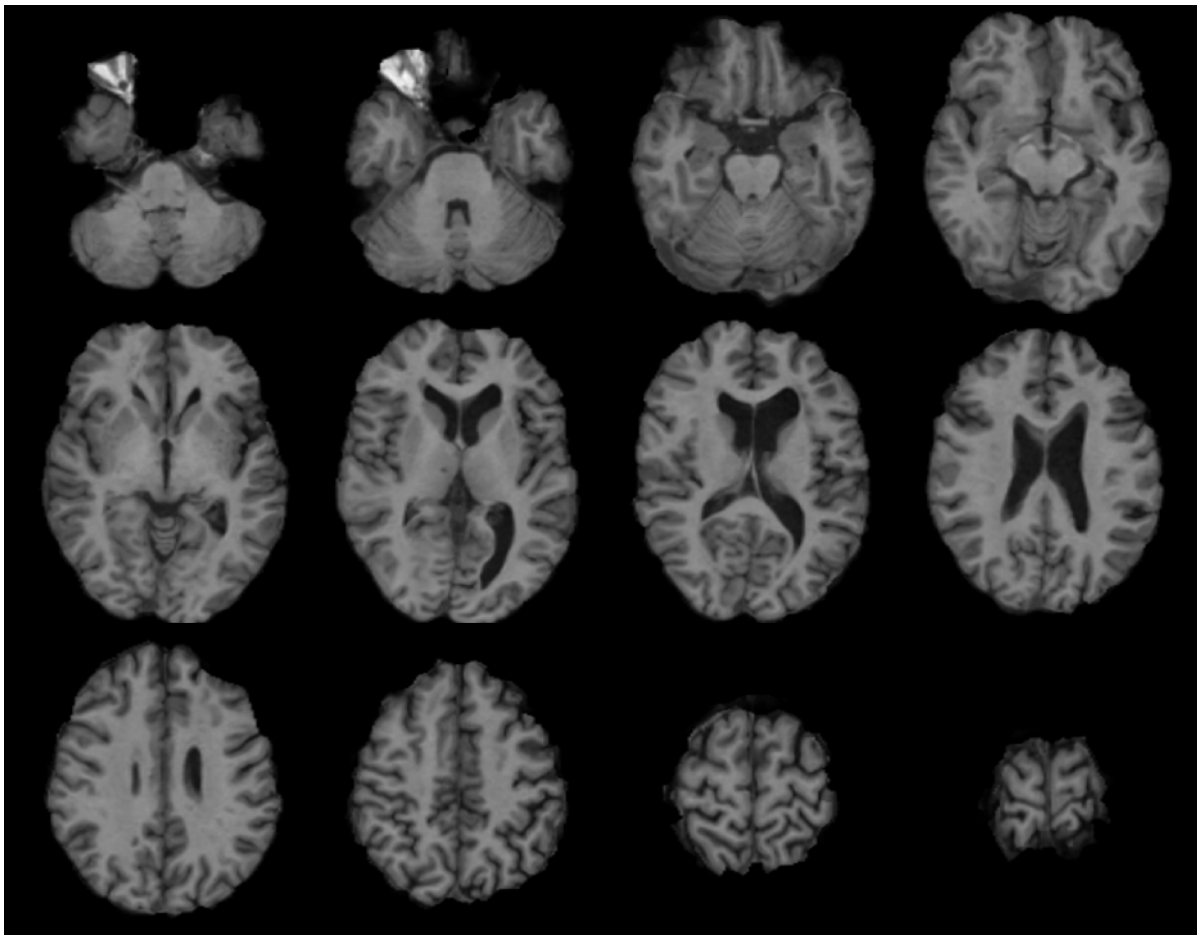


Figure 10: Example of good quality skull-stripped case with an error in the first two slice, in fact in the first two images is clearly evident the presence of one eye.

Table 6: Comparison of the QC of the skull stripping obtained after corrupting the ground truth with dilation and erosion in both spaces MNI and native. The table shows the mean probabilities of belonging to the positive class (good skull stripping) of some individual cases extracted from IBSR and MICCAI datasets.

		QUALITY CONTROL TEST			
		NATIVE SPACE		MNI SPACE	
Name dataset	**Case name**	**Eroded ground truth**	**Dilated ground truth**	**Eroded ground truth**	**Dilated ground truth**
IBSR	07	0.80	0.28	0.21	0.12
	08	0.76	0.35	0.25	0.21
	09	0.84	0.27	0.20	0.20
	11	0.78	0.20	0.23	0.15
MICCAI	01016SACH	0.74	0.40	0.15	0.11
	01038PAGU	0.84	0.35	0.20	0.18
	08027SYBR	0.85	0.23	0.18	0.20
	08029IVDI	0.73	0.32	0.25	0.23

index of the quality of the skull stripping. As a last step of the pipeline, we check if the QC index is higher than a given threshold; if not, the Brain Extraction procedure is repeated by modifying the algorithm (FSL-BET or ROBEX) parameters. The aim was to build an iterative process that uses, as a metric, the index given by the CNN to tune the possible parameters of the different tools; hence we use the quality index, computed from our module, to build a sort of feedback loop. Figure 13 shows the probability variation of 4 different MRI cases to be good skull stripping according to the variation of the parameters f. The parameter f was spanned between 0.2 and 0.8 with a step of 0.1, and the probability was computed with the second quality control algorithm developed QC in the MNI space.

From the graph is possible to understand that the stan-

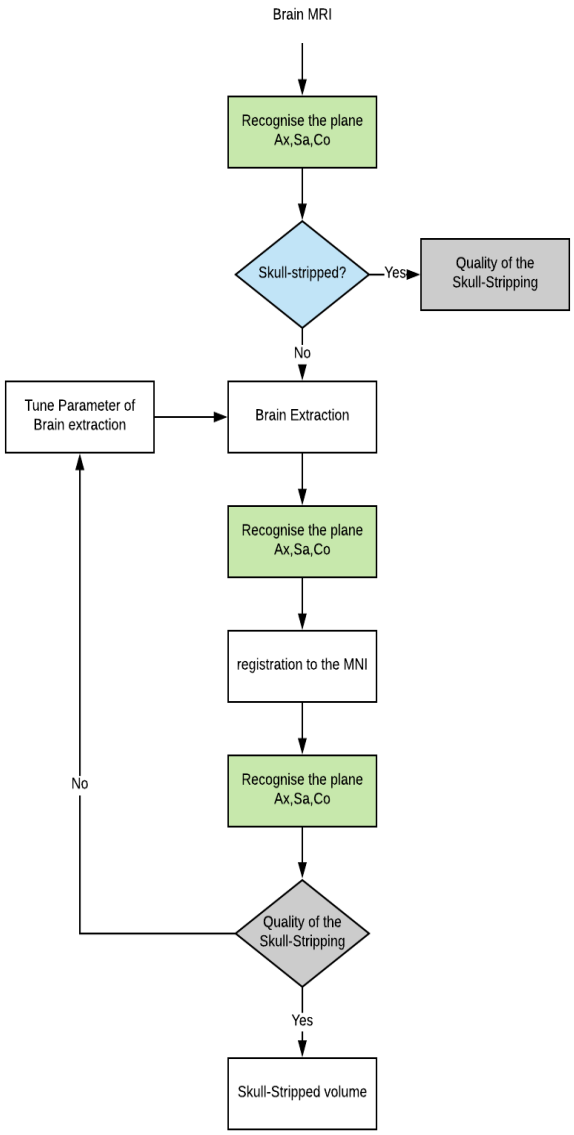


Figure 11: Example of a pipeline for analysing the quality of the skull stripping

dard value of the f parameter ($f = 0.5$) has the big mode. Hence, is the parameter that is more stable over all the cases. However, if we consider each case individually it can be seen that is not the best for everyone. For example, in case 1 and case 4 of the graph the best probability is reached with $f = 0.6$. After this, we visually inspect the skull-stripped volume to ensure the reliability of our quality control method.
Figure 12 shows the case 4 after applying on it FSL-BET with different values of the parameter f. The images respect totally what the graph shows; in fact when $f = 0.5$ the probability to be good is estimated around 0.79 and for $f = 0.6$ the probability is 0.86. In effect, by visual analysing the two result was possible to see the presence of one eye in the one with the lowest probability (see Figure 12).

5. Conclusions

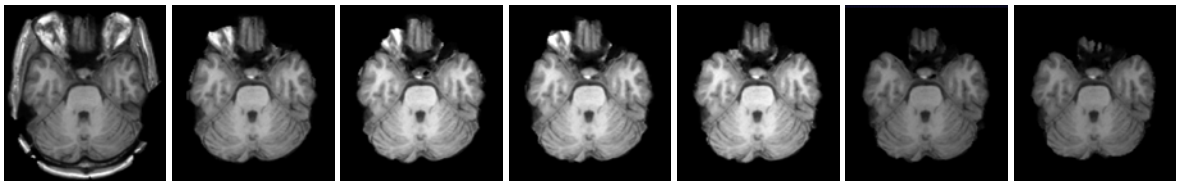
The evaluation of the quality of the medical images is not an easy task for different reasons such as the objectivity of the result or the absence of the ground truth. In this master thesis, we proposed and developed a fully automated deep learning quality assessment framework for online brain MRI processing. The prefixed objectives were achieved and more specifically, the work developed to reach the goal can be summarised as follows.

Transfer learning approach. Through the use of deep learning, we were able to solve many problems faced in this thesis such as the lack of labelled data or domain adaptation. When large training data is scarce, such as in medical imaging problems, a deep learning technique known as transfer learning has been demonstrated that is very effective (Ravishankar et al., 2017). For medical image problems, transfer learning is additionally attractive due to the heterogeneity of data types (modalities, anatomies, etc.) and clinical challenges. In conclusion about deep learning, we can say that it was interesting to see that a model, learnt for an unrelated problem setting, was able to solve a problem at hand with minimal retraining.

Data. For solving the problem of the domain adaptation we relied upon the adaptability of the pre-trained CNN. In fact, all the methods developed were able to work with cases taken from different domains. In order to prove this, we used nine different datasets during the test phase. Moreover, is important to point out that the datasets used in the test were totally new for the CNN.

Ground truth generation. Another important part of this thesis has been dedicated to generating synthetic labelled data. In order to achieve this, we mainly worked with morphology techniques. Through the use of these techniques, we were able to corrupt the ground truth in a realistic way. It is really important to generate a realistic corruption in order to allow the CNN to recognise similar cases during the classification. Moreover, by corrupting the data, we were able to test the model on real data taken from different domains.

Experimental. We developed mainly three modules in this master thesis: 1) plane recognition, 2) brain extraction recognition 3) quality control of the skull stripping. In all of them, we used a new approach based on morphology in order in order to select the central slice index of the volume and extract the most informative slices instead of using



(a) f=0.2 (b) f=0.3 (c) f=0.4 (d) f=0.5 (e) f=0.6 (f) f=0.7 (g) f=0.8

Figure 12: Skull stripping of the same volume and different values of the parameter f of the FSL-BET tool.

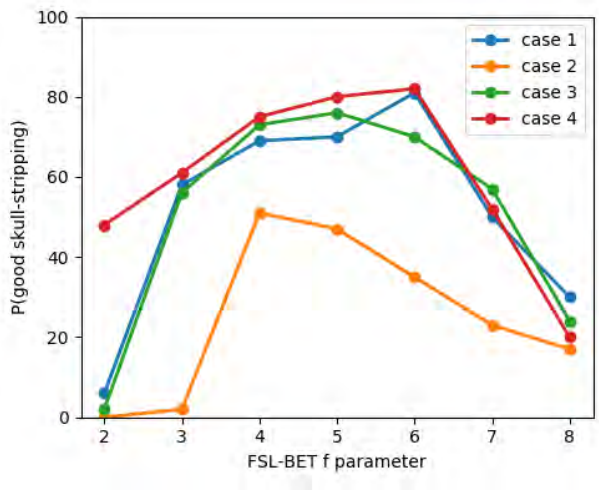


Figure 13: BET parameter selection showing the outcome probability of accurate skull stripping for different autotuned f values.

the whole volume for fine-tuning a CNN.

Platform integration. Regarding the platform integration, we successfully installed the web platform XNAT. Moreover, we built a container through the use of NVIDIA docker in order to plug it into the XNAT.

Special problems were faced in the module dedicated to the skull stripping quality. In particular, the CNN was not able to capture when the skull stripping was excessively eroded. For this reason, we developed a novel approach based on a metric computed between the volume analysed and the MNI mean volume. However, many diseases can influence the volumes of the brain. This can be problematic for our approach. Therefore, as a future work, we can think to crop the ventricles (that is the structure mainly affected by the atrophy) from the volume and analyse the rest. Really interesting is the possibility to use the different modules for building different pipelines. In fact, as an example, we built a pipeline able to autotune a parameter for the FSL-BET tool. Another step that should be investigate in more detail is the connection between the container and XNAT.

6. Acknowledgments

Firstly, I would like to express my gratitude to my supervisors Dr Sergi Valverde, Dr Arnau Oliver and Dr Xavier Lladó for the fundamental role they have played in this master thesis, which would not have been possible without their help. They are exceptional persons, both professionally and humanly speaking.
I would also like to thank the professors of the University of Bourgogne and the University of Cassino. A special gratitude goes to Prof. Francesco Tortorella who has always supported me.
I would like give a shout-out to Albert and Mostafa for making my work and my days in Girona easier. Last but not least I wish the best to all my friends in MAIA. In particular to my friends Luca, Ezequiel, Carmen, Sharon and Benjamin for sharing with me many great moments.
Most importantly, I want to thanks my family. Without them all of this would not be possible.
No, Milena do not worry, I could never forget about you. Thank you so much for the support for loving me and for the biscuits sent to the LAB. Your endless support always gives me the strength to carry on.
Finally, I want to say to the multiple sclerosis, thank you for giving me the passion for the medical field. You're the reason I'm here. I will try to win this battle with all my might.

References

Abadi, M., et al., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. URL: `https://www.tensorflow.org/`. software available from tensorflow.org.

Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D.L., Erickson, B.J., 2017. Deep learning for brain mri segmentation: state of the art and future directions. Journal of digital imaging 30, 449–459.

Albers, G.W., Thijs, V.N., Wechsler, L., Kemp, S., Schlaug, G., Skalabrin, E., Bammer, R., Kakuda, W., Lansberg, M.G., Shuaib, A., et al., 2006. Magnetic resonance imaging profiles predict clinical response to early reperfusion: the diffusion and perfusion imaging evaluation for understanding stroke evolution (defuse) study. Annals of neurology 60, 508–517.

Ardekani, B.A., Bachman, A.H., 2009. Model-based automatic detection of the anterior and posterior commissures on mri scans. Neuroimage 46, 677–682.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W., 2010. A theory of learning from different domains. Machine learning 79, 151–175.

Bennett, C.M., Miller, M.B., 2010. How reliable are the results from functional magnetic resonance imaging? Annals of the New York Academy of Sciences 1191, 133–155.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard, M.A., Rambaut, A., Drummond, A.J., 2014. Beast 2: a software platform for bayesian evolutionary analysis. PLoS computational biology 10, e1003537.

Cerqueira, M.D., Weissman, N.J., Dilsizian, V., Jacobs, A.K., Kaul, S., Laskey, W.K., Pennell, D.J., Rumberger, J.A., Ryan, T., Verani, M.S., et al., 2002. Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart: a statement for healthcare professionals from the cardiac imaging committee of the council on clinical cardiology of the american heart association. Circulation 105, 539–542.

Chabat, F., Hansell, D.M., Yang, G.Z., 2000. Computerized decision support in medical imaging. IEEE Engineering in medicine and Biology Magazine 19, 89–96.

Cheng, J.Z., Ni, D., Chou, Y.H., Qin, J., Tiu, C.M., Chang, Y.C., Huang, C.S., Shen, D., Chen, C.M., 2016. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in ct scans. Scientific reports 6, 24454.

Chollet, F., et al., 2015. Keras. `https://keras.io`.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. ImageNet: A Large-Scale Hierarchical Image Database, in: CVPR09.

Friedman, L., Glover, G.H., 2006. Report on a multicenter fmri quality assurance protocol. Journal of Magnetic Resonance Imaging 23, 827–839.

Gardner, E.A., Ellis, J.H., Hyde, R.J., Aisen, A.M., Quint, D.J., Carson, P.L., 1995. Detection of degradation of magnetic resonance (mr) images: comparison of an automated mr image-quality analysis system with trained human observers. Academic radiology 2, 277–281.

Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., et al., 2013. The minimal preprocessing pipelines for the human connectome project. Neuroimage 80, 105–124.

Glover, G.H., Mueller, B.A., Turner, J.A., Van Erp, T.G., Liu, T.T., Greve, D.N., Voyvodic, J.T., Rasmussen, J., Brown, G.G., Keator, D.B., et al., 2012. Function biomedical informatics research network recommendations for prospective multicenter functional mri studies. Journal of Magnetic Resonance Imaging 36, 39–54.

Herrick, R., Horton, W., Olsen, T., McKay, M., Archie, K.A., Marcus, D.S., 2016. Xnat central: Open sourcing imaging research data. Neuroimage 124, 1093–1096.

Iglesias, J.E., Liu, C.Y., Thompson, P.M., Tu, Z., 2011. Robust brain extraction across datasets and comparison with publicly available methods. IEEE transactions on medical imaging 30, 1617–1634.

Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L Whitwell, J., Ward, C., et al., 2008. The alzheimer's disease neuroimaging initiative (adni): Mri methods. Journal of magnetic resonance imaging 27.

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.

Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M., 2012. Fsl. Neuroimage 62, 782–790.

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Medical image analysis 5, 143–156.

Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P., 2016. On large-batch training for deep learning: Generalization gap and sharp minima. arXiv preprint .

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Kooi, T., Litjens, G., van Ginneken, B., Gubern-Mérida, A., Sánchez, C.I., Mann, R., den Heeten, A., Karssemeijer, N., 2017. Large scale deep learning for computer aided detection of mammographic lesions. Medical image analysis 35, 303–312.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. nature 521.

Li, T., 2012. Contributions to Mean Shift filtering and segmentation : Application to MRI ischemic data. Theses. INSA de Lyon. URL: `https://tel.archives-ouvertes.fr/tel-00768315`.

Liew, S.L., Anglin, J.M., Banks, N.W., Sondag, M., Ito, K.L., Kim, H., Chan, J., Ito, J., Jung, C., Khoshab, N., et al., 2018. A large, open source dataset of stroke anatomical brain images and manual lesion segmentations. Scientific data 5, 180011.

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.

Madan, C.R., Kensinger, E.A., 2016. Cortical complexity as a measure of age-related brain atrophy. NeuroImage 134, 617–629.

Osteaux, M., et al., 1992. A second generation pacs concept. Springer-Verlag .

Ravishankar, H., Sudhakar, P., Venkataramani, R., Thiruvenkadam, S., Annangi, P., Babu, N., Vaidya, V., 2017. Understanding the mechanisms of deep transfer learning for medical images. arXiv preprint arXiv:1704.06040 .

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, pp. 91–99.

Simonyan, K., Zisserman, A., 2014a. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556.

Simonyan, K., Zisserman, A., 2014b. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Smith, S.M., 2002. Fast robust automated brain extraction. Human brain mapping 17, 143–155.

Souza, R., Lucena, O., Garrafa, J., Gobbi, D., Saluzzi, M., Appenzeller, S., Rittner, R., Lotufo, R., 2017. An open, multi-vendor, multi-field-strength brain mr dataset and analysis of publicly available skull stripping methods agreement. NeuroImage .

Strother, S.C., 2006. Evaluating fmri preprocessing pipelines. IEEE Engineering in Medicine and Biology Magazine 25, 27–41.

Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE transactions on medical imaging 35, 1299–1312.

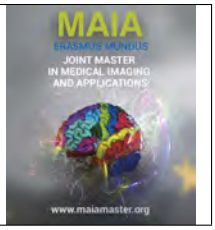
Vasilakos, A.V., Tang, Y., Yao, Y., et al., 2016. Neural networks for computer-aided diagnosis in medicine: A review. Neurocomputing 216, 700–708.

Warach, S., Dashe, J.F., Edelman, R.R., 1996. Clinical outcome in ischemic stroke predicted by early diffusion-weighted and perfusion magnetic resonance imaging: a preliminary analysis. Journal of Cerebral Blood Flow & Metabolism 16, 53–59.

Zhang, T., Koutsouleris, N., Meisenzahl, E., Davatzikos, C., 2014. Heterogeneity of structural brain changes in subtypes of schizophrenia revealed using magnetic resonance imaging pattern analysis. Schizophrenia bulletin 41, 74–84.

Medical Imaging and Applications

Master Thesis, June 2018



Prediction of Occult Invasive Disease in Ductal Carcinoma in Situ through Transfer Learning and Fine Tuning

Usama Pervaiz, MAIA team

Erasmus Mundus Joint Master Degree in Medical Imaging and Applications
University of Girona, Spain; University of Cassino, Italy; University of Burgundy, France

Joseph Y. Lo

Duke Univ. School of Medicine (United States)

Abstract

Purpose: The aim of the study is to determine whether the deep features extracted from the mammogram images using a convolutional neural network (ConvNet) are prognostic of occult invasive disease in ductal carcinoma in situ (DCIS). The major potential benefit of proposed tool can be to help evaluate the likelihood of patients in developing invasive ductal carcinoma (IDC).

Methods: In this study, we used the Duke University Medical Center data of Full Field Digital Mammograms (FFDM) of 340 unique patients. The deep convolutional is pre-trained on non-medical images (e.g, buildings, insects, animals) and then fine-tuned on indirectly related mammogram images. The ConvNet is fine-tuned using 65 IDC and 135 Atypical Ductal Hyperplasia (ADH) cases respectively and then system is independently tested on 35 ductal carcinoma with occult invasion (DCIS upstaged) and 105 pure DCIS images. The designed tool can output quantitative score which directly correlates with the severity and prevalence of DCIS.

Results: The deep features were able to distinguish DCIS upstaged cases from pure DCIS cases on patient level testing on a test data set with an Area under the Curve (AUC) of 0.72.

Conclusion: The performance outperformed the existing methods which used hand crafted traditional Computer Vision (CV) features. The proposed tool using the deep features extracted from mammogram images can perform significant role as a biomarker to monitor the DCIS progression.

Keywords: Ductal Carcinoma, Deep Learning, Fine Tuning, Convolutional Neural Nework, DCIS, Atypical Ductal hyperplasia, DCIS upstaged, digital mammogram, Breast cancer

1. Introduction

Ductal means that the cancer starts inside the milk ducts, carcinoma refers to any cancer that begins in the skin or other tissues (including breast tissue) that cover or line the internal organs, and in situ means "in its original place". Ductal carcinoma in situ (DCIS) is defined as existence of abnormal cells inside a milk duct in the breast. DCIS is known as one of the earliest form of breast cancer. DCIS is noninvasive, meaning it is contained within the milk duct and hasn't invaded other parts of the breast.

According to the American Cancer Society, about 60,000 cases of DCIS are diagnosed in the United States each year, accounting for about 1 out of every 5 new breast cancer cases. In recent years, because of the large scale use of mammography, the incidence of DCIS

has increased to sixfold. For the diagnosis of DCIS, we mostly rely on the detection of mammographically significant micro-calcifications (Holland and Hendriks, 1994). Although, we should be aware of the unusual radiographic manifestations of this disease; Ikeda and Andersson (1989) retrospectively analyzed the mammograms of 190 women with biopsy-proved DCIS and have shown 30 (16%) had negative mammograms, and 43 (23%) had mammographic manifestations of breast malignancy other than micro-calcifications. There are number of DCIS features linked with more aggressive biology have been identified, including high nuclear grade and comedonecrosis, although, many other features are a matter of further investigation (Mascaro et al., 2010).

DCIS is considered normally as noninvasive and not life-threatening, although if abnormal cells grow beyond ducts and gland, DCIS may progress into invasive cancer later on. Among biopsy-proven DCIS patients, approximately 20-56% are upstaged to reveal invasive ductal carcinoma (IDC) at the time of definitive surgery, while pure DCIS patients are not. Chin-Lenn et al. (2014) reported that out of 148 patients who underwent total mastectomy (TM), upstaging to invasive cancer at surgery occurred in 23%. Szynglarewicz et al. (2015) presented in this study that sixty-three women with pure DCIS presenting as sonographic mass lesion underwent core-needle biopsy and 56% of DCIS were upstaged. There can be number of factors associated associated with upstaging of DCIS. Studies using multivariate analysis found that a palpable lesion, a lesion size > 20 mm, a high grade lesion, radiological factors (BI-RADS category), factors related to CNB technique (modality of image guidance, size of the core needle, number of cores), were independently associated with upstaging of DCIS (Wiratkapun et al., 2011) (Kim et al., 2012).

To understand the concept of DCIS grading, we need to first revisit different prior stages which leads to DCIS. Hyperplasia also known as epithelial hyperplasia or proliferative breast disease is an overgrowth of cells that line the ducts or milk glands (lobules). As shown in Fig 1, Ductal Hyperplasia (DH) refers to overgrowth of cells lining milk ducts. Atypical ductal hyperplasia (ADH) has increased abnormal growth pattern of cells, although it is not considered as breast cancer. Rather, it is a marker that there is risk for breast cancer in the future. DCIS cases are regarded as stage-0 cancer, and the cancer cells are still limited within milk ducts. Studies of loss of heterozygosity in low-grade DCIS and ADH have revealed similar genetic changes in the two conditions (Lakhani et al., 1995); this finding confirmed that these are clonal processes and both fulfill the basic concept of neoplasia (Pinder and Ellis, 2003a). Invasive ductal carcinoma (IDC), sometimes called infiltrating ductal carcinoma, refers to cancer that has surpassed the wall of milk duct and is invading other breast tissues.

In this paper, we will refer DCIS as *negative* cases, upstaged DCIS as *positive* cases. We will use ADH cases as negative class during training, so we will refer them as *super-negatives* and IDC cases will be used as positive class during training, so they will be referred as *super-positives*. We will target to distinguish pure DCIS from upstaged DCIS cases. This is a clinically challenging task with a high inter and intra reader variabilities. It is worth noticing that 10 to 44% of DCIS patients will go through re-operation to evaluate their regional lymph nodes because of previously undetected IDC (Cox et al., 2001). The ability to predict occult invasion can avoid delays in definitive diagnosis and can reduced a significant amount of cost (Cox et al., 2001).

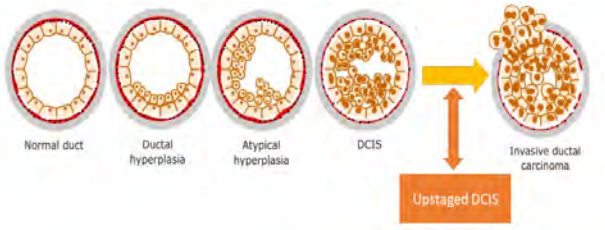


Figure 1: **Different Pathology classes of DCIS**, Image modified from (Orenstein, 2014)

The aims of this study were as follows:

1. Propose a potential noninvasive replacement technique for the traditional invasive methods of grading DCIS.
2. Address the classification of pure DCIS and upstaged cases using mammograms in conjunction with deep learning with artificial neural networks.
3. Demonstrate a baseline application for using ConvNet in DCIS upstaging and efficiency.

2. State of the art

In this section, firstly we have discussed the clinical correlation of DCIS and DCIS upstaged cases and then we segued to actual models.

There are a number of techniques in literature targeting histological features and medical imaging analysis, proposed to predict DCIS upstaging. It is of paramount importance to realize that about one in four DCIS diagnoses at core-needle biopsy (CNB) represent understaged invasive breast cancer. Brennan et al. (2011) concluded that preoperative variables significantly associated with under-staging include high-grade lesion at CNB (vs non high grade lesion, P < .001), lesion size larger than 20 mm at imaging

(vs lesions ≤ 20 mm, P < .001), and mammographic mass (vs calcification only, P < .001). Bagnall et al. (2001) presented some clinical features associated with DCIS upstaging; high grade core biopsy DCIS and >40 calcifications translate to 48% invasive at surgical histology; high grade core biopsy DCIS and <40 calcifications translate to 15% invasive; non-high grade core biopsy DCIS resulted in 0% invasive). Dillon et al. (2006) have shown that some of mammographic features including size ≥ 5 cm on excision pathology was linked with higher risk of invasion (P = 0.002). Lee et al. (2016) demonstrated that immunohistochemical evidence of human epidermal growth factor receptor 2 overexpression (P=0.010) was also predictive of DCIS upstaging. O'Flynn et al. (2009) have determined that there is significant associations with the presence of invasive disease for cluster size (p=0.0001) and DCIS grade (p=0.003). Park et al. (2013) constructed a nomogram and predicted the likelihood of DCIS invasive cancer with an AUC of 0.71.

On the other hand, Lee et al. (2000) concluded that mammographic and histologic features cannot be used reliably to predict cases that are underestimated with stereotactic core needle biopsy (SCNB). Renshaw (2002) also stated that neither the radiographic findings, presence of comedonecrosis, comedo histology, lobular extension, size of the largest focus, nor aggregate size was significantly associated with an increased incidence of invasion. The papers' above shows that medical imaging findings, immunohistochemical evidences' and histological features have very limited power in predicting DCIS upstaging. Also, most of these techniques relies on invasive methods.

An increasing number of medical imaging techniques which align with computer-based classification and segmentation algorithms are also being examined and validated by researchers. These analytic methodologies are being applied to different types of medical images for DCIS staging. For the same speicific task as we studied, three have been three previous papers all from our institution. Shi et al. (2018a) shows that ConvNet pre-trained on only non-medical images, extracted deep features were able to distinguish DCIS with occult invasion from pure DCIS, with an AUC of 0.68. Using the handcrafted 113 mammographic features, the multivariate classifier was able to distinguish DCIS with occult invasion from pure DCIS, with an AUC of 0.70 (Shi et al., 2017a). Zhu et al. (2017) revealed that algorithmically assessed MRI features predict DCIS upstaging with an AUC=0.719.

There are a number of CAD schemes proposed in literature but none of them have been physically implemented in any clinical practice by now for the detection or screening of the DCIS upstaging. Hence there is an utmost need for a CAD tool, which can not only distinguish between DCIS cases and DCIS upstaged cases, but also can predict the disease quantification on a mammogram. The task of DCIS upstaging analysis is very challenging with the use of traditional machine learning algorithms, since there are no well-defined characteristics of the DCIS or DCIS upstaged abnormal cells. In recent years, ConvNets have rapidly emerged as a widespread machine learning technique in a number of applications especially in the area of medical image classification and segmentation. This problem inspired us to choose ConvNet as the deep artificial neural network (ANN) for our analysis.

3. Materials

We collected the Full Field Digital Mammograms (FFDM) from 340 unique patients at Duke University Medical Center with Institutional Review Board (IRB) approval. Those patients magnification views of mammograms were also acquired by a GE Senographe Essential FFDM system, with a magnification factor of 1.5 or 1.8 times. The data was categorized into three groups; ADH, DCIS and IDC. We also excluded some cases based on the exclusion criteria which is the presence of any masses, asymmetries, or architectural distortion in a mammogram; history of breast cancer or prior surgery; and presence of microinvasion at the time of initial biopsy.

Specifically from those 340 patients mammogramic images, 135 of cases were ADH, 140 of cases were DCIS, and 65 of cases were IDC. DCIS patients underwent stereotactic core needle biopsy (SCNB) and were diagnosed with DCIS prior to surgical removal of the tumor. Among them, 35 cases were found as invasive later during surgery (either lumpectomy or mastectomy), and the rest 105 cases were not. The region of interest (ROI) mask for the ADH, DCIS and IDC lesion in each subject was delineated by an expert breast radiologist. We also collected magnification views of 85 healthy subjects which are free from any calcifications or masses. We further extracted 7792 patches from these magnification views.

Fig 2 shows an example of segmentation result of individual microcalcifications and detection of cluster boundary.

4. Methods

Before we go into details of methodology, the overall process is summarized in a form of flowchart in Fig 3. It shows that first step is collection of data in Duke Department of Radiology, and then micro-calcifcations

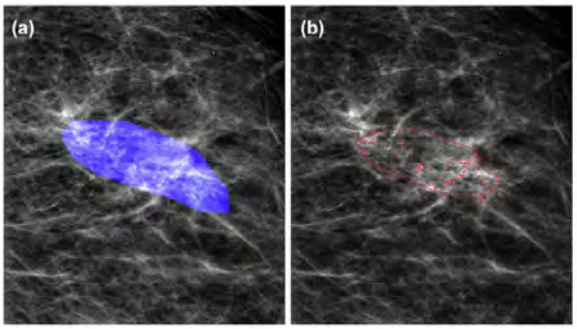


Figure 2: **A: Segmented by algorithm, B: delineated by a radiologist, Image taken from (Shi et al., 2017b)**

were delineated by a radiologist. After that, we generated the Region of Interest's (ROIs) from the algorithm. The ConvNet is trained on ROIs of ADH and IDC and then tested on ROIs of DCIS and DCIS upstaged cases. In the end, we analyzed the results both quantitatively and qualitatively.

4.1. Pre-Processing

This section will cover the pre-processing techniques applied on the magnification views of mammogram images.

4.1.1. Paddle Detection

Most of the magnification views contains a bright boundary (paddle), which should not be included when you extract patches. The first step is detection of rough ROI from DICOM images excluding those paddles. The algorithm is based on noise tolerant peak finding algorithm (Yoder, 2011), Gaussian model fitting (Zivkovic, 2004) and subsequently probability distribution of image (Parzen, 1962).

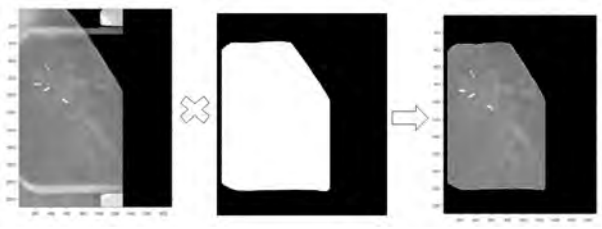


Figure 4: **Mask generation from Dicom images**

4.1.2. Patch Extraction

The second step is the extraction of patch containing the breast micro-calcifications by drawing rectangular bounding box around the ROI manually delineated by radiologist. Before feeding any image to the system, pixel intensities of the images were re-scaled to [0,0.9] and, contrast of the input image is improved by applying contrasted limited adaptive histogram equalization (Clipping limit= 0.005)(Zuiderveld, 1994) and gamma correction (gamma value = 0.5) (Farid, 2001). The processed patch is shown at an extreme right in Fig 5.

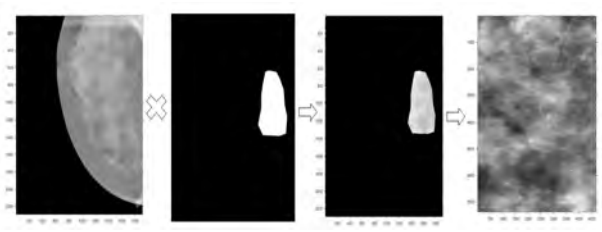


Figure 5: **Region of Interest Patch Extraction**

4.1.3. Patch Rescaling

The patches extracted from ADH, IDC or DCIS upstaged cases are of varying size as shown in Fig 6, where 40 cases of invasive breast cancer (IDC) is shown as an example. Most of the convolutional neural networks required a fix image size, that's why images were rescaled to a fixed dimension. In our study, if the input image size is NxM, where N>224 or M>224, then we took square ROIs of [224x224], randomly shifted to at least include 80% of the cluster mask. On the other hand, if N<224 or M<224, we padded image with zeros to rescaled it to [224x224]. The other way around which we attempted is to interpolate the patch to [224x224] but this caused distortions, changing the imaging attributes like texture, density of tissues associated with the calcifications and surrounding tissues.

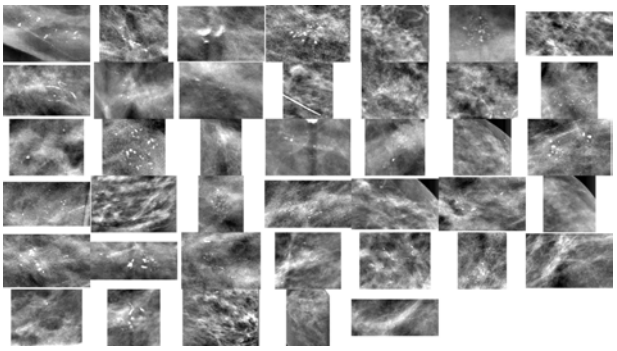


Figure 6: **Sample 40 IDC cases**

4.1.4. Patch Neighborhood

In the Bilinear Neural Network (BNN), we would be taking into account global context by considering a much larger area surrounding calcifications, based on the hypothesis that meaningful information lies beyond the local neighborhood of calcifications.

We considered the fixed image of dimension [448*448]. If the size of patch is (N,M), and N<448 and M<448, we extracted a larger patch of [448*448] based on the centered of smaller region of interest (ROI), by using 100% pixels only as displayed in Fig 7. If N>448 or M>448, we first centered that patch on original image and down-sampled ROI to [448*448].

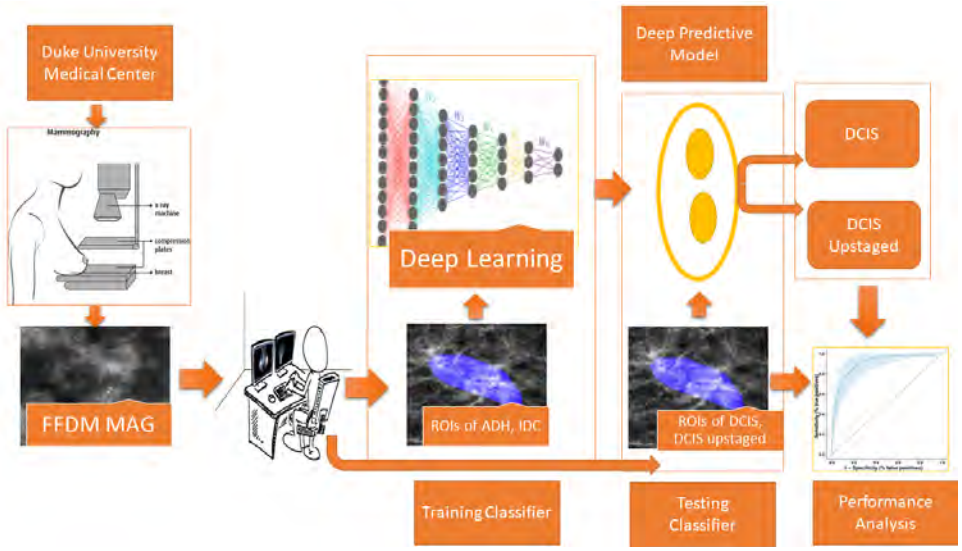


Figure 3: **The overall methodology of presented CAD tool**

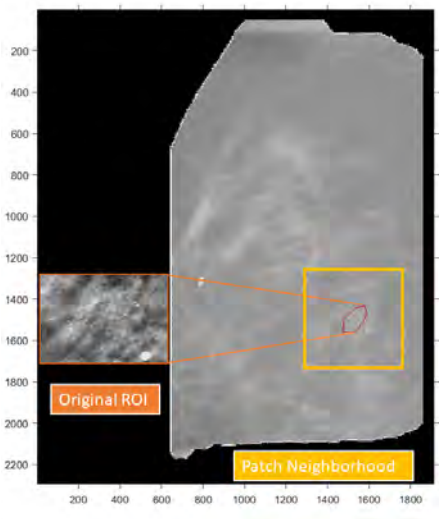


Figure 7: **Local neighborhood of calcifications**

4.1.5. Data Augmentation

The data augmentation was performed real time in the data space to improve the performance of the classifier and to avoid over-fitting. Table 1 shows the various data augmentation techniques applied only on the training data.

4.2. Training without Transfer Learning

We trained a three convolutional layer neural network (ConvNet-3) and a four layer convolutional layer neural network (ConvNet-4) using the super positives (IDC cases) and super negatives (ADH cases). Table 2 shows detailed architecture of (ConvNet-3) and (ConvNet-4). The input image of [224*224] passes through a number of convolution and pooling layers. In both of the architectures, two fully connected layers were used with number of neurons as 512 and 256 respectively. The hyper parameter details are as following; stochastic gradient descent was used as an optimizer, categorical cross entropy as loss function, sigmoid as an activation function, batch size of 32, learning rate of 0.0001 and momentum of 0.9. The final model architecture was selected based on the maximum performance from the validation data set after a number of experiments.

4.3. Training with Transfer Learning and Fine-tuning

4.3.1. Network Architecture

The transfer learning and fine-tuning approach applied in the proposed architecture which can be explained as follows. The VGG16 convolutional neural architecture was designed and then pre-trained ImageNet model weights were loaded into the system. The fully-connected layers were removed, then the rest of the ConvNet was treated as a feature extractor for the new dataset. Once the features for all images were extracted, a classifier was attached and trained for the new dataset. In the end, weights of the pre-trained network were fine-tuned via backpropagation by un-freezing the lower convolutional layers and retraining more layers. We tried different combinations of frozen and fine tuned layers. The final proposed architecture approach is shown in Fig 8.

The ConvNet used is an modified version of the VGG16 model, with a 15 layer weight structure, (13 convolutional (conv) layers and 2 fully connected (fc) layers) (Page et al., 1982). We used the VGG16 architecture because it has small (3x3) convolution filters and depth of 15 layers, which efficiently captures extremely minute details. The input volume is [224*224*3], passed through a stack of convolutional layers where

Table 1: **Data Augmentation**

Arguments	Parameters	Comments
Rotation Range	30°	Random Rotations from -15° to + 15°
Width Shift Range	0.2	Range for random horizontal shifts (fraction of total width)
Width Height Range	0.2	Range for random vertical shift (fraction of total width)
Shear Range	0.2	Shear Intensity (Shear angle in counter-clockwise direction as radians)
Zoom Range	0.2	Range for random zoom.
Horizontal Flip	True	Randomly flip inputs horizontally

Table 2: **Comparing Network Architectures, filter number * filter size (e.g., 64 * 3^2), filter stride (e.g., str 2), pooling window size (e.g., pool 2^2,), and the output feature map size (e.g., map size 112 *112)**

Model	$conv_1$	$conv_2$	$conv_3$	$conv_4$
ConvNet-3	64*3^2, str 2	128*3^2, str 2	256*3^2, str 2	
Modified	pool 2^2, str2	pool 2^2, str2		-
	map size 112*112	56*56	28*28	
ConvNet-4	64*3^2, str 2	128*3^2, str 2	256*3^2, str 2	512*3^2, str 2
Modified	pool 2^2, str2	pool 2^2, str2	pool 2^2, str2	-
	map size 112*112	56*56	28*28	14*14

we used the filters with a small receptive field of [3*3]. The first convolution layer computes the output connected with a local region and results in a volume of [224*224*64] where 64 is the number of filters. Then, there is a max pooling layer which reduces' the size of each patch by half, for example [224*224*64] will be down-sampled to [122*122*64]. In total there are five max-pooling layers with a stride value of 2. Moreover, all hidden layers are also equipped with rectification non-linearity (ReLU) (Castro et al., 2008). The conv. layers are followed with two fully connected layers with 512 and 256 channels respectively. In the end, sigmoid layer uses an activation function, giving an output between 0 to 1 for binary classification. The detailed architecture of VGG16 model is shown in Fig 9.

4.3.2. Network Training and Parameters

The ConvNet was trained using the training and validation data set. The number of samples that were propagated along ConvNet for each iteration, known as batch size, was set to 32. The entire set of data passed through the whole network for 50 times, after that there was no substantial increase in performance for further epochs. The network was trained at a small learning rate of 0.00001 to make sure that updates magnitudes were kept minimal for our transfer learning approach. We used a considerably large value of 0.9 for momentum, in order to ensure that the network doesn't get stuck in local minima during fine tuning. The most optimum results were obtained using the RMSProp optimizer for training the bottleneck features, and applying Stochastic gradient descent optimizer for fine tuning. The binary cross-entropy was applied as a loss function, and sigmoid as an activation function. For performance evaluation metrics, accuracy, precision and recall were implemented. The check point for the training was the accuracy on the validation set so the training will stop automatically once the accuracy on validation set stopped increasing. The data is split into training and validation(80%,20%). The final model architecture was selected based on the maximum performance from the validation data set.

4.3.3. Comparative Analysis with other Pretrained Network architectures

There can be an argument in favor of using an architecture with a lesser number of layers instead of using the VGGG16 like AlexNet (Krizhevsky et al., 2012), or a more deep convolutional neural network like ResNet (He et al., 2016a) for pre-training. To be certain, we performed a comparative analysis between VGG16, AlexNet and ResNet.

Fig 10 illustrates the design of our proposed AlexNet, including the different layers in the ConvNet architecture. The input image passes through different sets of layers. These layers consist of the convolution layer, max-pooling layer and rectified linear unit (ReLU) layer. Moving to the network's rear part, the architecture includes a fully-connected layer and a sigmoid loss layer, which ensures that the output of the network represents class to which the input image belongs. The batch size was set to 32, stochastic gradient descent was used as an optimizer. The binary cross-entropy was applied as a loss function, and accuracy was used for performance evaluation.

The comparative network architecture of VGG16 and AlexNet is summarized in Table 3 and deep ResNet ar-

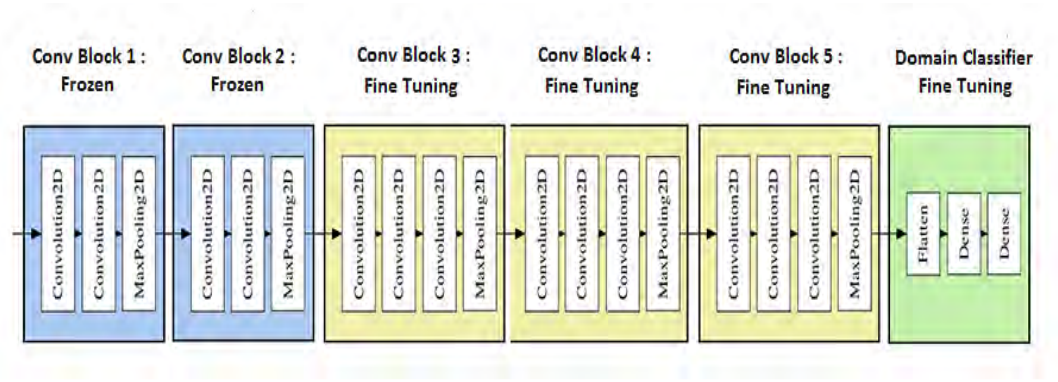


Figure 8: **The layer structure of proposed architecture.**
The first two convolution (conv) blocks were kept frozen, the conv blocks depicted in yellow color were fine tuned and lastly, the green block consists of of the fully connected classification layers. It should be noted that for simplicity, we didn't show the max-pooling layers.

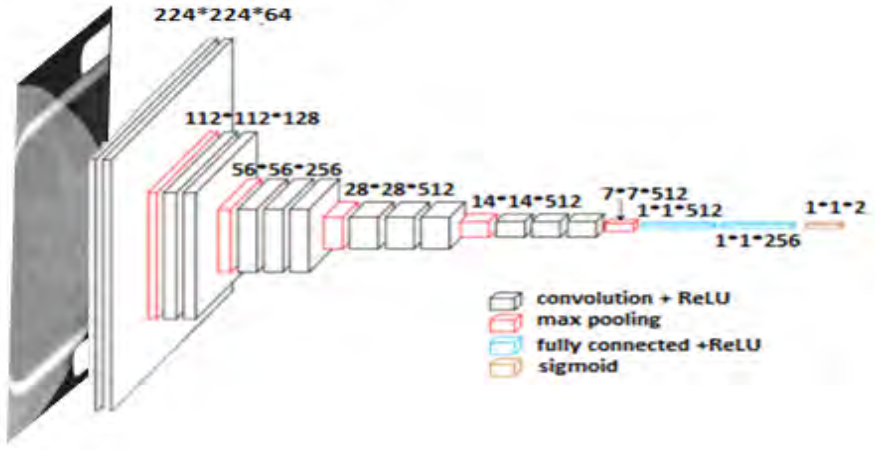


Figure 9: **Detailed Network Architecture,** Modified from (Khawaldeh et al., 2017)
The input image is passed through a set of convolution, pooling layers and fully connected layers.

chitecture can be referred in (He et al., 2016a). We used an 18 layer ResNet, with five covolution blocks, fixed kernel size of [3*3], stride of 2, 1000 neurons in the fully connected layer and sigmoid as an activation function. The floating point operations per second is $1.8 * 10^9$ and near 0.2 million parameters.

4.4. Bilinear Neural Network

We propose bilinear models, a recognition architecture that consists of two feature extractors whose outputs are pooled to obtain an image descriptor. There are two different input streams, the first channel takes ROI patch as input, and second channel takes surrounding neighborhood of ROI patch. This second channel will keep into account of any meaningful and relevant information present in encompassing region of patch consequently utilizing the global context.

4.4.1. Training without Transfer Learning

In this experiment, we randomly initialize the weights of ConvNet following a gaussian distribution and trained from scratched on super positives (IDC) and super negatives (ADH) images. We designed a two channel network here which is not based on any existing architecture. The first channel inputs an image of [224*224] and second channel inputs an image of [448*448]. We kept the number of filters and kernel size to be small to avoid overfitting. The first channel has 15488 parameters and second channel has 80000 parameters. There are two fully connected dense layers for each network and we extract 256 deep features from each channel. We concatenate the 256 feature vector from each layer and then used a drop out rate of 0.5. In the end, there are two more fully connected layers which are pulling the features from both channel and outputting a continuous score between 0 to 1 for each case using sigmoid activation function. We also have used Batch Normalization (BatchNorm) (Ioffe and Szegedy, 2015), a widely adopted technique that enables faster and more stable training of deep neural networks (DNNs), also it makes the optimization landscape significantly smoother. The architecture is presented in

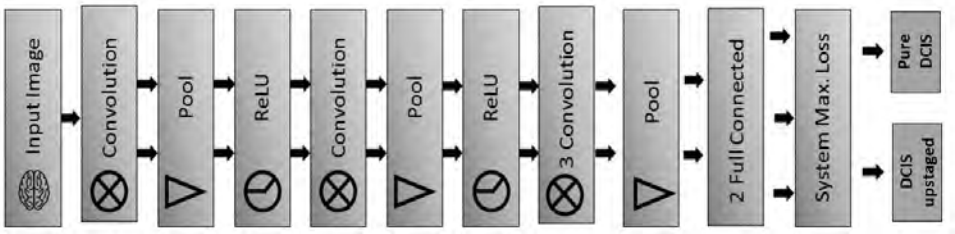


Figure 10: **Layer Structure of AlexNet**

Table 3: **The network parameters of proposed ConvNets**

Model	$conv_1^2$	$conv_2^2$	$conv_3^3$	$conv_4^3$	$conv_5^3$
VGG16	64*3^2, str 2 pool 2^2, str2 map size 112*112	128*3^2, str 2 pool 2^2, str2 56*56	256*3^2, str 2 pool 2^2, str2 28*28	512*3^2, str 2 pool 2^2, str2 14*14	512*3^2, str 2 pool 2^2, str2 7*7
AlexNet	96*11^2, str 2 pool 3^2, str2 map size 55*55	128*5^2, str 2 pool 3^2, str2 27*27	384*3^2, str 2 13*13	384*3^2, str 2 13*13	256*3^2, str 2 13*13

$conv_1^2$ (The first conv block with two identical convolution layers), filter number * filter size (e.g., 64 * 3^2), filter stride (e.g., str 2), pooling window size (e.g., pool 3^2,), and the output feature map size (e.g., map size 112 *112)

Appendix 1 in Fig 22. For consistency, we will call it BNN1 through out the paper.

4.4.2. Training with Transfer Learning and Fine-tuning

In this experiment, we didn't randomly initialize the weights of ConvNet but used pre-trained weights from ImageNet. There are also two different feature extractors based on pre-trained convolutional neural networks. The CNN stream A is based on VGG16 and CNN stream B is based on ResNet, and both networks are initialized from the ImageNet dataset followed by domain specific fine-tuning. The entire model is fine tuned using back-propagation for several epochs (about 45 to 50) and a fairly small learning rate of (l = 0.00001).The fully-connected layer is invariably positioned as the last part of the ConvNet architecture and is responsible for assigning class scores in supervised settings and sigmoid loss layer is used for the performance evaluation for each input. This architecture as shown in Fig 11 is inspired from Lin et al. (2015), however it significantly differs as in Lin et al. (2015) , they used a single input channel. For consistency, we will call it BNN2 through out the paper.

4.5. Performance Evaluation

In medical imaging analysis, particularity, for the binary classification problem, it is important to note that positive indicates the disease is present and negative indicates the disease is absent. In our case, Sensitivity (positive for disease) refers to the proportion of subjects who have the upstaged DCIS (reference standard positive). Alternatively, Specificity (negativity in health) is the proportion of subjects without the upstaged DCIS and give negative test results.These performance evaluation measures are shown in Eq (1) and Eq (2) .

$$Sensitivity/Recall = \frac{TP}{TP + FN} \tag{1}$$

$$Specificty = \frac{TN}{FP + TN} \tag{2}$$

The sensitivity (Sen) and specificity (Spe) vary across the different threshold and the sensitivity is inversely related with specificity. Then, the plot of sensitivity versus 1-Specificity is called receiver operating characteristic (ROC) curve and the area under the curve (AUC), as an effective measure of accuracy.

4.6. Class Activation Maps (CAM)

Class activation maps are a simple technique to get the discriminative image regions used by a CNN to identify a specific class in the image. To create the CAM, we restricted the network to have a global average pooling layer after last convolutional layer. After that, we added a dense layer (softmax activation). We can't directly apply this technique to existing networks like VGG16 as they have fully connected layers. To make it work, we modified existing networks and then fine-tuned on domain specific data.

The most important block of CAM is global average pooling (GAP). We have a N dimensional image after last convolutional layer like in VGG16 we have N=512 number of filters in the last convolutional layer. If we discard fully connected layer, we can use input image

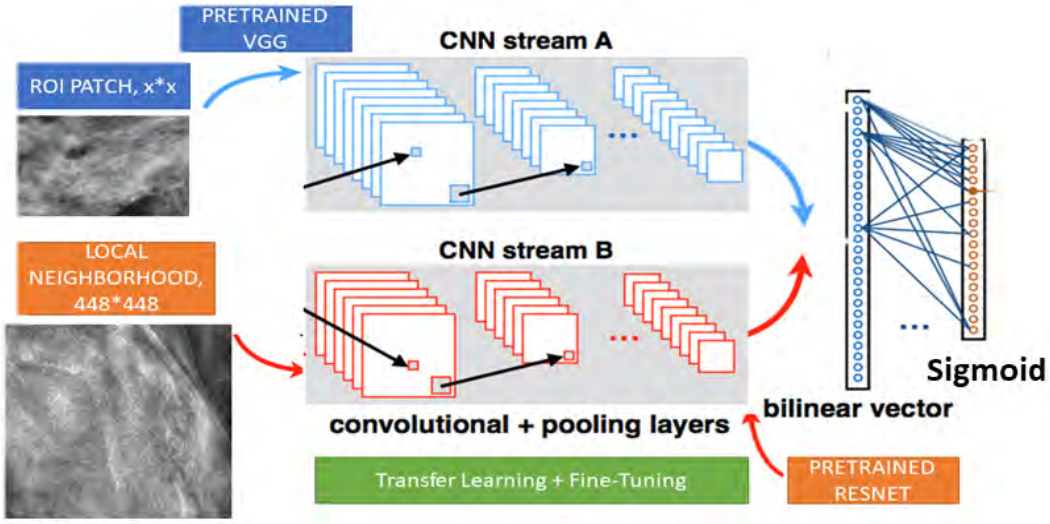


Figure 11: **Bilinear Neural Network Architecture.**
The input image is passed through a set of convolution, pooling layers and fully connected layers.

of any size, for example if our input image has a resolution of [512*512], the output shape of last convolutional layer will be 512*64*64. We know that, 512/64 =8, our spatial resolution mapping will be [8*8]. GAP will take all 512 channels and spatial average will be returned. Channels which have high activations will result in high signals. The final output will be heatmap for every convolutional layer. The CAM is also known sometimes as saliency maps (Li and Yu, 2015).

5. Results

5.1. Quantitative Results

For training and validation of models', we used ADH, IDC and negative patches. We divided data into different configurations for our negative and positive class as shown in Table 4. Data was partitioned into training and validation by a 80% and 20% split respectively.

As explained in methodology, we trained different ConvNet architectures. In table 5, we briefly summarized these experiment settings. When there is no transfer learning, it means that network was trained from scratch on of the data configuration from Table 4. When there is transfer learning, it means that we used pretrained weights of ImageNet model. When there is fine-tuning on top of transfer learning, it meant that we re-trained model last few layers on domain targeted data set (C1,C2 or C3). We tried with different combinations of freezing and fine-tuning of convolutional layers. The most optimum configuration for VGG16 was; freezing the first two convolution blocks and fine tuning on last three convolution blocks, for AlexNet was; freezing the first two convolutional layers and fine tuning on last three convolution layers, for ResNet18; freezing the first four convolution blocks (12 convolutional layers) and fine tuning on last convolution block (four convolutional layers with 512 kernels in each layer).

We also proposed two different Bilinear Neural Network(BNN) architectures which have not been presented in any literature before. BNN1 is the configuration when we trained model only on domain targeted dataset (C1,C2 or C3) without any transfer learning or fine tuning. BNN2 is the configuration, when we used pre-trained ImageNet weights for models (VGG16 and ResNet18). For simplicity, we named each model as with a combination of α and β symbols, as shown in Table 5. DCIS and DCIS upstaged data was not exposed to network anytime during training and validation. This will ensure that there is no over-fitting and we only used pure DCIS and DCIS upstaged cases for testing of the models. The results shown in Table 6 is on the test dataset of 105 DCIS cases and 35 DCIS upstaged cases. It shows the AUC for various experiment and data configuration following Table 4 and Table 5.

Fig 12 shows that fine-tuning of VGG16 architecture significantly improve result on test dataset from AUC of 0.58 to 0.72. Fig 13 shows the ROC curves for pre-trained and fine-tuned model of AlexNet (AUC= 0.70) and ResNet (AUC=0.72). Fig 14a shows the ROC curve for BNN1 and, Fig 14b shows ROC curve for BNN2, and demonstrates how transfer learning improved AUC to 0.68.

Table 4: **Data Configuration**

Experiment Name	Negative Class	Positive Class
C1	ADH	IDC
C2	Negative Patches and ADH	IDC
C3	Negative Patches	ADH and IDC

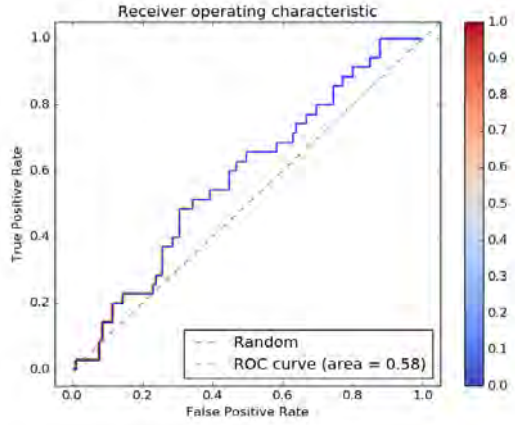
Table 5: **Experiment Configuration : Transfer Learning (TL), Fine-Tuning (FT), Mammograms Data (ADH, IDC and Negative Patches)**

Experiment Name	ConvNet Model	Model Config	Model Details
Exp $\alpha 1\,\beta 1$	ConvNet-3	TL = No, FT= No	Trained only on Mammograms
Exp $\alpha 1\,\beta 2$	ConvNet-4	TL = No, FT= No	Trained only on Mammograms
Exp $\alpha 2\,\beta 1$	VGG16	TL= Yes, FT= No	Trained only on ImageNet
Exp $\alpha 2\,\beta 2$	AlexNet	TL= Yes, FT= No	Trained only on ImageNet
Exp $\alpha 2\,\beta 3$	ResNet	TL = Yes, FT= No	Trained only on ImageNet
Exp $\alpha 3\,\beta 1$	VGG16	TL = Yes, FT= Yes	Trained on ImageNet & Mammograms
Exp $\alpha 3\,\beta 2$	AlexNet	TL = Yes, FT= Yes	Trained on ImageNet & Mammograms
Exp $\alpha 3\,\beta 3$	ResNet	TL = Yes, FT= Yes	Trained on ImageNet & Mammograms
Exp $\alpha 4\,\beta 1$	BNN1	TL = No, FT= No	Trained only on Mammograms
Exp $\alpha 4\,\beta 2$	BNN2	TL = Yes, FT= No	Trained only on ImageNet
Exp $\alpha 4\,\beta 3$	BNN2	TL= Yes, FT= Yes	Trained on ImageNet & Mammograms

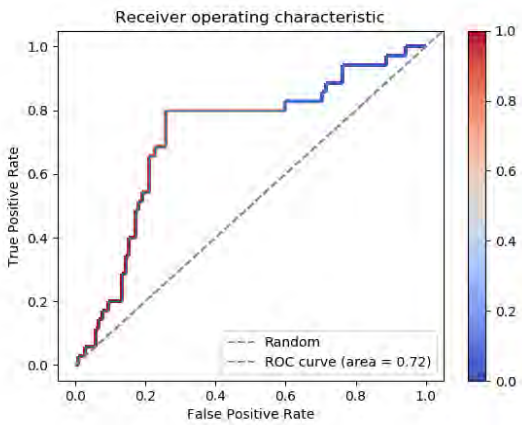
Table 6: **AUCs for different Experiment Configurations**

Data Configurations	**C1**	**C2**	**C3**
Experiment Configurations	**AUC**	**AUC**	**AUC**
Exp $\alpha 1\,\beta 1$	.55	0.56	0.51
Exp $\alpha 1\,\beta 2$	0.58	0.58	0.50
Exp $\alpha 2\,\beta 1$		0.58	
Exp $\alpha 2\,\beta 2$		0.61	
Exp $\alpha 2\,\beta 3$		0.60	
Exp $\alpha 3\,\beta 1$	0.70	0.72	0.56
Exp $\alpha 3\,\beta 2$	0.71	0.72	0.58
Exp $\alpha 3\,\beta 3$	-	0.70	0.57
Exp $\alpha 4\,\beta 1$	0.64	0.65	0.59
Exp $\alpha 4\,\beta 2$		0.63	
Exp $\alpha 4\,\beta 3$	0.68	0.68	0.59

Table 7: AUCs for different Experiment Configurations

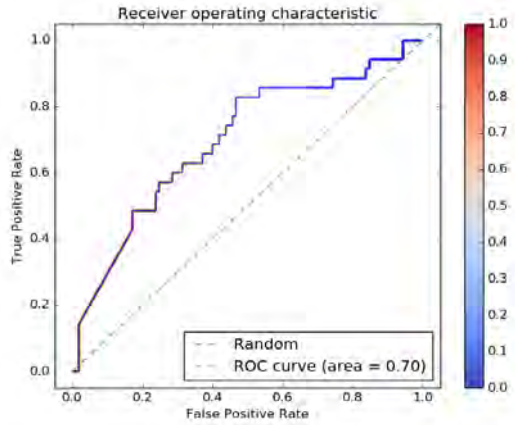


(a) AUC Of Exp $\alpha 2\,\beta 1$ C2

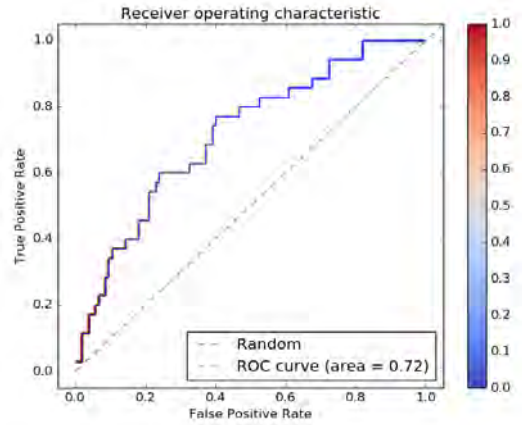


(b) AUC Of Exp $\alpha 3\,\beta 1$ C2

Figure 12: AUC of VGG16 before and after fine-tuning

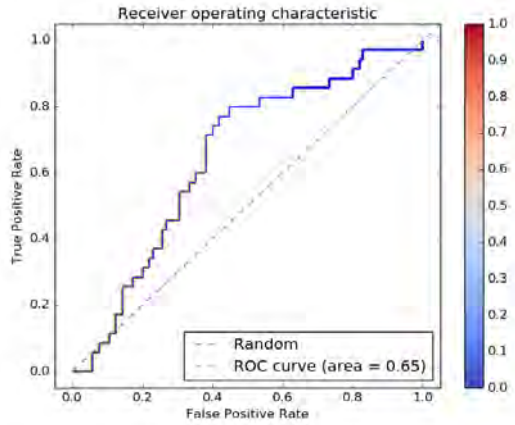


(a) AUC Of Exp $\alpha 3\,\beta 2$ C2

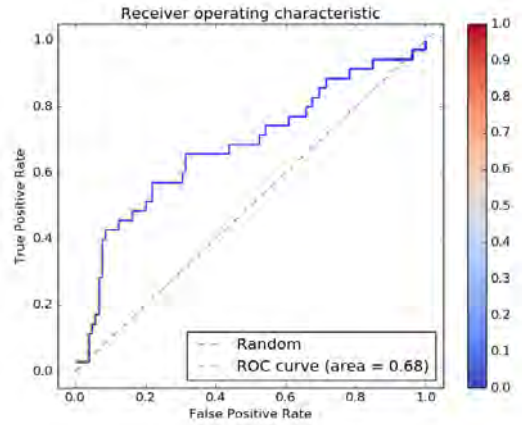


(b) AUC Of Exp $\alpha 3\,\beta 3$ C2

Figure 13: AUC of ResNet and AlexNet after fine-tuning



(a) AUC Of Exp $\alpha 4\,\beta 1$ C2



(b) AUC Of Exp $\alpha 4\,\beta 3$ C2

Figure 14: AUC of Bilinear Neural Network without and with Transfer Learning

Fig15 shows the clustered column graph for only C2 data configuration for Exp $\alpha 2$, Exp $\alpha 3$ and Exp $\alpha 4$. It can be seen that transfer learning and fine-tuning subsequently helped to improve the performance.

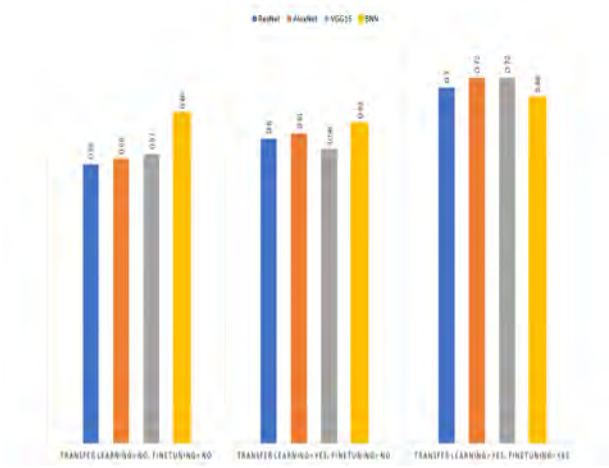


Figure 15: **Comparative Performance of Different Architectures**

Fig16 shows the clustered column graph for C1, C2 and C3 data configuration for Exp $\alpha 3$, Exp $\alpha 4$. It can be seen that C2 is best data configuration, very comparable to C1 but both of them are significantly better than C3.

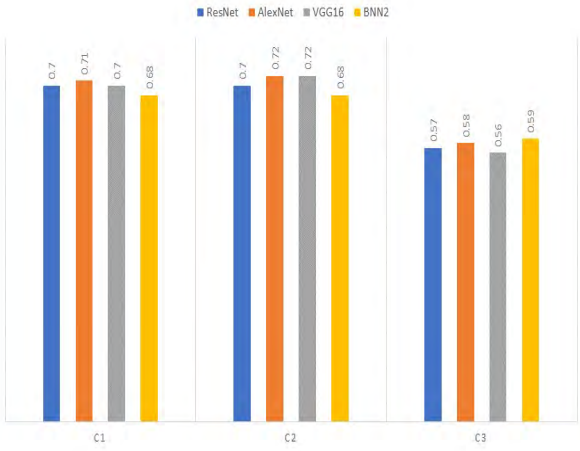


Figure 16: **Comparative Performance of Different Data Configurations**

5.2. *Qualitative Results*

CAM lets us see which regions in the image were relevant to this class. Zhou et al. (2016) shows that CAM allows re-using classifiers for getting good localization results. This will actually show what network is learning and which parts of image are activated or play significant role when exposed to model for testing.

As we have seen in previous section that fine-tuned VGG16 outperforms the other presented convolutional neural network architectures, so we presented CAM only on VGG16 model.

5.2.1. *Pretrained VGG16 without Fine-tuning*

In Fig 17, we displayed saliency maps for VGG16 model pre-trained on image net weights (No Fine-tuning). Fig 17 shows saliency maps of two sample cases, it can be seen that most of activations are random. (Only 2 cases are shown for sake of simplicity, remaining data followed the same trend).

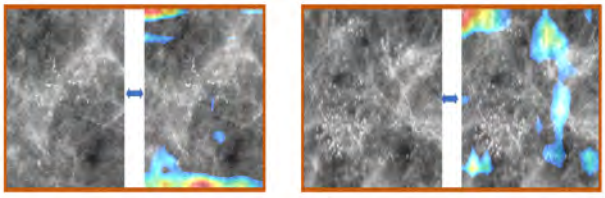


Figure 17: **Class Activation Maps on sample DCIS cases**

5.2.2. *Pretrained VGG16 with Fine-tuning*

In this case, we fine-tuned VGG16 on ADH and IDC cases. The saliency maps are shown in Fig 18, Fig 19, Fig 20 and Fig 21 of few true positives, false positives, true negatives, and false negative cases respectively. We further discuss these saliency maps in 6.

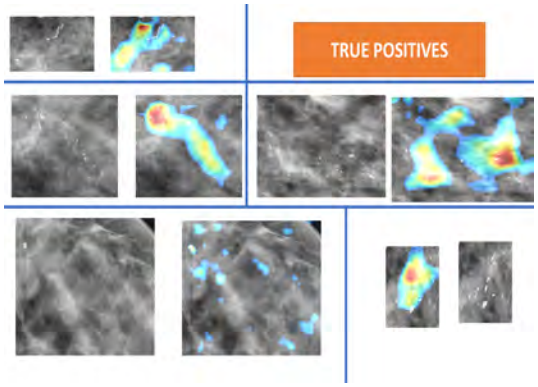


Figure 18: Saliency Maps of few True Positive Cases

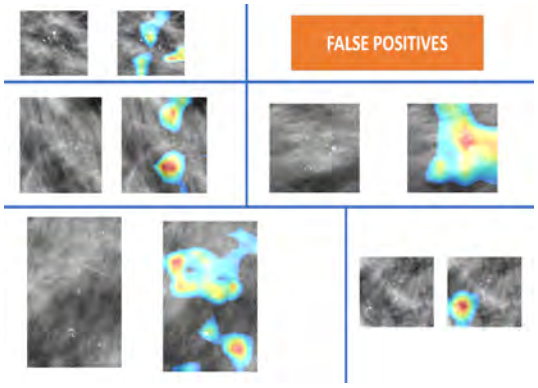


Figure 19: Saliency Maps of few False Positive Cases

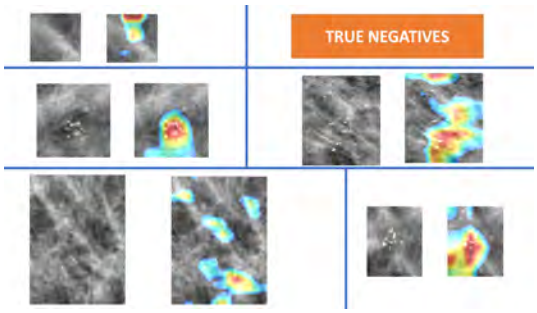


Figure 20: Saliency Maps of few True Negative Cases

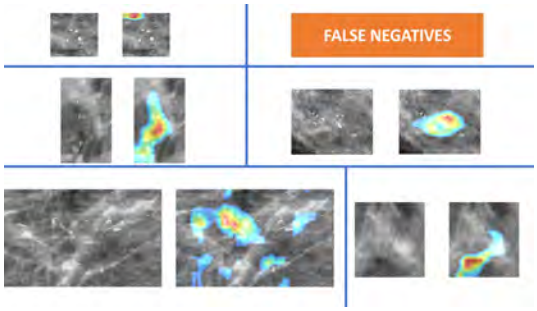


Figure 21: Saliency Maps of few False Negative Cases

6. Discussion

Our presented study aims to tackle the challenging DCIS upstaging problem with a lot of unique aspects. We fine-tuned a ConvNet using a set of super positives and super negatives images, which are not directly related to targeted problem, to have a generic target representation. The reason we have not directly used the DCIS cases is because of data limitation as we only have 105 DCIS and 35 DCIS upstaged cases. We didn't expose DCIS data to network during training and validation, hence it was reserved for testing only so it can serve as an independent test data set.

There may be an argument that unrelated data for training might not help to learn useful features for DCIS upstaging prediction. As briefly explained in 1, DCIS and ADH shared some similar overlapping features and have revealed similar genetic changes in two diseases; It is often interpreted as confirmatory evidence that they are clonal processes and both therefore fulfill the basic concept of neoplasia. There is a strong argument that this boundary between ADH and DCIS is not ideal on a morphological, immunohistochemical, or genetic basis (Pinder and Ellis, 2003b). Based on this, we used ADH as negative class for training and DCIS as negative class for testing as they share common generic features, and ConvNet would still be able to learn differentiating features.

On the other hand, we used IDC as a positive class during training and DCIS upstaged cases as positive class while testing. He et al. (2016b) demonstrated DCIS as a true precursor of IDC is indirect but convincing: IDC is rarely seen without adjacent DCIS (Wong et al., 2010); DCIS and IDC from the same patient share similar genetic features Page et al. (1982) and molecular abnormalities (Castro et al., 2008). Also, upstaged DCIS is literally a mix of DCIS and IDC. Based on these findings, it make sense to indirectly used IDC for fine-tuning to predict DCIS upstaging.

We also have demonstrated in this paper, the importance of transfer learning, by first training the neural network on non-domain specific images like animals and spiders. ConvNet-3, ConvNet-4 and BNN1 are trained directly on the breast cancer images shows inferior results. Hence, we first transfer pre-trained weights from ImageNet model (trained on non-medical images) which reasonably improve the classification performance. We also trained VGG16 and AlexNet from scratch on breast cancer images but models' couldn't perform well mainly because of limited data. This also highlights the significance of transfer learning when you are facing the challenge of insufficient data for medical imaging analysis.

Transfer learning is helpful but it is also shows an inadequate and restricted performance for this challenging task. Recently, neural networks are proposed which consist of hundreds of hidden layers and million of parameters. The lower layer of network learns only the basic features like shape of edges, blobs which can be learn from non-medical images. The deep layers of network learns the features related to final goal. Therefore, it is crucial to wisely fine-tune ConvNet, we demonstrated that fine-tuning last few layers with a smaller learning rate on mammogram images significantly increase the performance as shown in Fig 12a and Fig 12b. However, fine-tuning more than few layers would cause no further improvement and model will overfit data.

The most optimum results are achieved using VGG16 network with an AUC of 0.72, shown in Fig 12b. The results through AlexNet model are also comparable with an AUC of 0.72 but with less sensitivity and specificity as shown in Fig 12b and Fig 13b. One of the reason VGG16 performed better than the other networks is very small convolutional filters of size (3*3). This small receptive field of VGG16 helped network to capture the minimal changes, like particularly appropriate for smaller calcifications, which can help ConvNet to differentiate DCIS from DCIS upstaged cases. We have shown in results that AUC is equal to 0.70 when using pre-trained ResNet. This proves wrong another common misconception that

deeper the network is, better the results are, but it is highly dependent on the amount of data and, when you have a limited data, it is better to shift to less deeper architecture.

This process of DCIS upstaging is tedious and an extremely challenging task for radiologist's and pathologist's and subject to inter- and intra-reader variabilities. Shi et al. (2017b) shows that histologic features including nuclear grade and DCIS subtype did not show statistically significant differences between cases with pure DCIS and with DCIS upstaged cases. Only three mammographic features, i.e., the major axis length of DCIS lesion, the BI-RADS level of suspicion, and radiologist's assessment showed reasonable results with an AUC-ROC equal to 0.62. In our work, we have shown that transfer learning using non-medical images and fine-tuning on indirectly related mammogram images have helped us to achieve an AUC of 0.72.

It is also important to comment on different data configurations results as shown in Fig 16. From our experiments, we have concluded that if we use negative patches in one class and both ADH an IDC cases in other class, then system performs badly as it learns only to detect the calcifications and it will treat both DCIS and DCIS upstaged cases as positive. The best data division was when we used ADH and few negative patches as negative class and IDC cases as positive class. This is an interesting finding as we realized how to balance training classes in most optimal way for calcification's detection versus classification task.

We also conducted numerous experiments based on hypothesis to consider global context of disease by considering tissues surrounding DCIS and upstaged DCIS abnormal cells. We trained numerous networks by using a combination of different bilinear architectures training in parallel like VGG16-VGG16 or VGG16-ResNet. Ganeshan et al. (2011) compared uniformity values in focal lesions and surrounding tissue and showed significant differences between DCIS with or without invasive carcinoma (IC) versus IC ($p = 0.0009$). This pilot showed the potential for computer-based assessments of heterogeneity within focal mammographic lesions and surrounding tissue to identify adverse pathological features in mammographic lesions. However, our experiments based on extracting the deep features by exploring the surrounding tissues of DCIS and DCIS upstaged cases resulted in negative results and a drop in performance as shown in Fig 14a and Fig 14b. One of the limitation we faced which can be a potential reason is limited memory of GPU. Training two deep architectures in parallel required a larger RAM, so we had to decrease batch size to 6 to cope with this issue. Another reason can be difference in surrounding tissues of DCIS or DCIS upstaged cases with ADH or IDC cases. We fine-tuned network on surrounding neighborhood of IDC and ADH cases. There is no evidence in literature which can proof that surrounding tissues of IDC correlates with that of DCIS upstaged or ADH surrounding tissues has same kind of heterogeneity as of DCIS cases outside the focal lesions.

We will briefly compare here results with existing solutions using computer vision (CV) and deep learning based classifiers as presented in 2. Our group has previously shown that using the handcrafted 113 mammographic features, the multivariate classifier was able to distinguish DCIS with occult invasion from pure DCIS, with an AUC of 0.70 (Shi et al., 2017a). However, this paper was based only on 99 cases and based on leave one out (LOO) cross validation. However, when our group using similar CV feautures moved to 140 DCIS cases, and reported the median AUC from repeated cross validation, resulting in AUC of only 0.61 (Hou et al., 2018). Shi et al. (2018a) shows that ConvNet pre-trained on non-medical images, extracted deep features were able to distinguish DCIS with occult invasion from pure DCIS, with an AUC of 0.68, however this result was also based on 99 cases and LOO cross validation. Further disadvantage of this technique is that feature response and feature selection are driven by DCIS and DCIS upstaged cases so there is a chance of over-fitting. Shi et al. (2018b) pre-trained a ConvNet on ImageNet and Inbreast dataset, and achieved an AUC of 0.75 but there are certain disadvantages with that approach also. Firstly, it performs the stability feature selection on deep learning based extracted features (One have to manually pull features from different layers and check features response to DCIS data) unlike our process in which we used the fully connected layers for features extraction. Our newly presented approach make the whole procedure more viable to be implemented in clinical practice as it is a single step, fully automatic approach. Specifically, since we used the pre-trained network on classification task, the extracted features were more suited for DCIS upstaging prediction. Moreover, Shi et al. (2018b) used DCIS data for training and validation (20% cases (7 positives, 21 negatives) were randomly selected for validation, while the remaining 80% (28 positives, 84 negatives) were used for training) which certainly posed the risk of over-fitting as compare to our approach which was trained on semi-related data and never exposed to DCIS and DCIS upstaged data.

We also have plotted the CAM of DCIS and DCIS upstaged cases. Although, it is visually very difficult to differentiate between DCIS and DCIS upstaged cases, hence challenging to comment on CAM for this classification problem. DCIS upstaged is related with heterogeneity; local recurrences are seen much more frequently with high grade DCIS (Lagios, 1995), and

it can be seen from activation maps of true positives that ConvNet is detecting this heterogeneity for DCIS upstaged cases. Also, these activation maps show that network is not trained on noise or basic computer vision features (like number or size of calcifications) but most of deep features are based on other factors like shape, texture, spread of calcification's and density changes around the focal ROI. However, we cannot validate any claims relating to CAM right now and this would need further work with radiologists and pathologists. The results presented here have lower AUC as normally we expect to see for medical imaging analysis problems. Our group has assigned the same task to radiologists' in past to classify DCIS and DCIS upstaged cases only based on mammogramic images and achieved an AUC of 0.69 based on one radiologist assessment, and other radiologist achieved an AUC of 0.5 only. This emphasize how hard this task is and demonstrate significance of our deep learning based extracted achieving an AUC of 0.72. Keeping in mind these things into consideration, our result can be considered as reliable and state of art for differentiating DCIS from DCIS upstaged cases.

One of the major potential benefit of this CAD tool can be to help evaluate the likelihood of patients in developing invasive breast cancer. As, numerous studies have shown the association between the DCIS upstaging and IDC. There are number of CAD schemes available to reliably detect breast calcification's or masses on mammograms. Our proposed tool can be incorporated with these existing CAD schemes, so when patients go for breast cancer screening, we can also examine for occurrence of DCIS or DCIS upstaging. Detection of DCIS at an early or mild stage can reduce the risk of developing invasive breast cancer and can notably ameliorate the breast function with proper intervention.

Our approach had certain limitations. Firstly, we are still trying to understand the real meaning behind the deep features that could also be interpretable by a clinician. Secondly, the size of test data set is small and results can not be generalizable. Thirdly, the AUC of 0.72 is not enough to be implemented in clinical ways so new ways should be sought out to improve the performance further. Lastly, we plan to combine the imaging based tool with pathology based tool which could further strengthen trust in designed classifier.

In summary, we have demonstrated the feasibility of employing a CAD scheme for prediction of occult invasive diseases in ductal carcinoma. Although, our results were promising and provide tools that may be implemented into practice, we still envision improvements for this tool in the future. For the future work, we will be testing our best performing model (VGG16 trained on ImageNet and fine-tuned on ADH and IDC data) on 138 more Duke cases, 115 of these cases are pure DCIS and 23 of them are upstaged cases and these cases are currently being annotated by a radiologist.

7. Conclusions

In conclusion, our study demonstrated the feasibility of using deep features for the radiomics DCIS upstaging prediction task. Specifically, we proposed a deep learning based model that was pre-trained on non-medical images, and fine-tuned on indirectly related mammogram images, applied to mammogram images for DCIS grading. From our work, we hope to show the medical imaging community, an alternate way of using the powerful deep learning technique when a large scale dataset is not available. With statistically significant results, as well as the ability to provide a quantitative index, this study may serve as first step in applying this tool in clinical settings. Future work will explore collaboration with other institutions to include more subjects to further validate the model.

8. Acknowledgments

I would like to acknowledge Rui Huo and Yinhao Ren who are part of Duke Radiology group and were part of weekly meetings. They helped me refine different ideas and also provided access to their GPUs when required.

References

Bagnall, M.J., Evans, A.J., Wilson, A.R.M., Pinder, S.E., Denley, H., Geraghty, J.G., Ellis, I.O., 2001. Predicting invasion in mammographically detected microcalcification. Clinical radiology 56, 828–832.

Brennan, M.E., Turner, R.M., Ciatto, S., Marinovich, M.L., French, J.R., Macaskill, P., Houssami, N., 2011. Ductal carcinoma in situ at core-needle biopsy: meta-analysis of underestimation and predictors of invasive breast cancer. Radiology 260, 119–128.

Castro, N.P., Osório, C.A., Torres, C., Bastos, E.P., Mourão-Neto, M., Soares, F.A., Brentani, H.P., Carraro, D.M., 2008. Evidence that molecular changes in cells occur before morphological alterations during the progression of breast ductal carcinoma. Breast Cancer Research 10, R87.

Chin-Lenn, L., Mack, L.A., Temple, W., Cherniak, W., Quinn, R.R., Ravani, P., Lewin, A.M., Quan, M.L., 2014. Predictors of treatment with mastectomy, use of sentinel lymph node biopsy and upstaging to invasive cancer in patients diagnosed with breast ductal carcinoma in situ (dcis) on core biopsy. Annals of surgical oncology 21, 66–73.

Cox, C.E., Nguyen, K., Gray, R.J., Salud, C., et al., 2001. Importance of lymphatic mapping in ductal carcinoma in situ (dcis): Why map dcis?/discussion. The American surgeon 67, 513.

Dillon, M.F., McDermott, E.W., Quinn, C.M., O'doherty, A., O'higgins, N., Hill, A.D., 2006. Predictors of invasive disease in breast cancer when core biopsy demonstrates dcis only. Journal of surgical oncology 93, 559–563.

Farid, H., 2001. Blind inverse gamma correction. IEEE Transactions on Image Processing 10, 1428–1433.

Ganeshan, B., Strukowska, O., Skogen, K., Young, R., Chatwin, C., Miles, K., 2011. Heterogeneity of focal breast lesions and surrounding tissue assessed by mammographic texture analysis: preliminary evidence of an association with tumor invasion and estrogen receptor status. Frontiers in oncology 1, 33.

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

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

Holland, R., Hendriks, J., 1994. Microcalcifications associated with ductal carcinoma in situ: mammographic-pathologic correlation., in: Seminars in diagnostic pathology, pp. 181–192.

Hou, R., Shi, B., Grimm, L.J., Mazurowski, M.A., Marks, J.R., King, L.M., Maley, C.C., Hwang, S., Lo, J.Y., 2018. Improving classification with forced labeling of other related classes: application to prediction of upstaged ductal carcinoma in situ using mammographic features, in: Medical Imaging 2018: Computer-Aided Diagnosis, International Society for Optics and Photonics. p. 105750R.

Ikeda, D., Andersson, I., 1989. Ductal carcinoma in situ: atypical mammographic appearances. Radiology 172, 661–666.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 .

Khawaldeh, S., Pervaiz, U., Rafiq, A., Alkhawaldeh, R.S., 2017. Noninvasive grading of glioma tumor using magnetic resonance imaging with convolutional neural networks. Applied Sciences 8, 27.

Kim, J., Han, W., Lee, J.W., You, J.M., Shin, H.C., Ahn, S.K., Moon, H.G., Cho, N., Moon, W.K., Park, I.a., et al., 2012. Factors associated with upstaging from ductal carcinoma in situ following core needle biopsy to invasive cancer in subsequent surgical excision. The Breast 21, 641–645.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

Lagios, M.D., 1995. Heterogeneity of duct carcinoma in situ (dcis): relationship of grade and subtype analysis to local recurrence and risk of invasive transformation. Cancer letters 90, 97–102.

Lakhani, S., Collins, N., Stratton, M., Sloane, J., 1995. Atypical ductal hyperplasia of the breast: clonal proliferation with loss of heterozygosity on chromosomes 16q and 17p. Journal of Clinical Pathology 48, 611–615.

Lee, C.H., Carter, D., Philpotts, L.E., Couce, M.E., Horvath, L.J., Lange, R.C., Tocino, I., 2000. Ductal carcinoma in situ diagnosed with stereotactic core needle biopsy: can invasion be predicted? Radiology 217, 466–470.

Lee, C.W., Wu, H.K., Lai, H.W., Wu, W.P., Chen, S.T., Chen, D.R., Chen, C.J., Kuo, S.J., 2016. Preoperative clinicopathologic factors and breast magnetic resonance imaging features can predict ductal carcinoma in situ with invasive components. European journal of radiology 85, 780–789.

Li, G., Yu, Y., 2015. Visual saliency based on multiscale deep features. arXiv preprint arXiv:1503.08663 .

Lin, T.Y., RoyChowdhury, A., Maji, S., 2015. Bilinear cnn models for fine-grained visual recognition, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1449–1457.

Mascaro, A., Farina, M., Gigli, R., Vitelli, C.E., Fortunato, L., 2010. Recent advances in the surgical care of breast cancer patients. World journal of surgical oncology 8, 5.

O'Flynn, E., Morel, J., Gonzalez, J., Dutt, N., Evans, D., Wasan, R., Michell, M., 2009. Prediction of the presence of invasive disease from the measurement of extent of malignant microcalcification on mammography and ductal carcinoma in situ grade at core biopsy. Clinical radiology 64, 178–183.

Orenstein, 2014. stage zero breast cancer. Internet:https://theriskybody.wordpress.com/tag/stage-zero-breast-cancer .

Page, D.L., Dupont, W.D., Rogers, L.W., Landenberger, M., 1982. Intraductal carcinoma of the breast: follow-up after biopsy only. Cancer 49, 751–758.

Park, H.S., Kim, H.Y., Park, S., Kim, E.K., Kim, S.I., Park, B.W., 2013. A nomogram for predicting underestimation of invasiveness in ductal carcinoma in situ diagnosed by preoperative needle biopsy. The Breast 22, 869–873.

Parzen, E., 1962. On estimation of a probability density function and mode. The annals of mathematical statistics 33, 1065–1076.

Pinder, S.E., Ellis, I.O., 2003a. The diagnosis and management of pre-invasive breast disease: ductal carcinoma in situ (dcis) and atypical ductal hyperplasia (adh)–current definitions and classification. Breast Cancer Research 5, 254.

Pinder, S.E., Ellis, I.O., 2003b. The diagnosis and management of pre-invasive breast disease: ductal carcinoma in situ (dcis) and atypical ductal hyperplasia (adh)–current definitions and classification. Breast Cancer Research 5, 254.

Renshaw, A.A., 2002. Predicting invasion in the excision specimen from breast core needle biopsy specimens with only ductal carcinoma in situ. Archives of pathology & laboratory medicine 126, 39–41.

Shi, B., Grimm, L.J., Mazurowski, M.A., Baker, J.A., Marks, J.R., King, L.M., Maley, C.C., Hwang, E.S., Lo, J.Y., 2017a. Can occult invasive disease in ductal carcinoma in situ be predicted using computer-extracted mammographic features? Academic radiology 24, 1139–1147.

Shi, B., Grimm, L.J., Mazurowski, M.A., Baker, J.A., Marks, J.R., King, L.M., Maley, C.C., Hwang, E.S., Lo, J.Y., 2018a. Prediction of occult invasive disease in ductal carcinoma in situ using deep learning features. Journal of the American College of Radiology 15, 527–534.

Shi, B., Grimm, L.J., Mazurowski, M.A., Marks, J.R., King, L.M., Maley, C.C., Hwang, E.S., Lo, J.Y., 2017b. Can upstaging of ductal carcinoma in situ be predicted at biopsy by histologic and mammographic features?, in: Medical Imaging 2017: Computer-Aided Diagnosis, International Society for Optics and Photonics. p. 101342X.

Shi, B., Hou, R., Mazurowski, M.A., Grimm, L.J., Ren, Y., Marks, J.R., King, L.M., Maley, C.C., Hwang, E.S., Lo, J.Y., 2018b. Learning better deep features for the prediction of occult invasive disease in ductal carcinoma in situ through transfer learning, in: Medical Imaging 2018: Computer-Aided Diagnosis, International Society for Optics and Photonics. p. 105752R.

Szynglarewicz, B., Kasprzak, P., Halon, A., Matkowski, R., 2015. Preoperatively diagnosed ductal cancers in situ of the breast presenting as even small masses are of high risk for the invasive cancer foci in postoperative specimen. World journal of surgical oncology 13, 218.

Wiratkapun, C., Patanajareet, P., Wibulpholprasert, B., Lertsithichai, P., 2011. Factors associated with upstaging of ductal carcinoma in situ diagnosed by core needle biopsy using imaging guidance. Japanese journal of radiology 29, 547.

Wong, H., Lau, S., Yau, T., Cheung, P., Epstein, R., 2010. Presence of an in situ component is associated with reduced biological aggressiveness of size-matched invasive breast cancer. British journal of cancer 102, 1391.

Yoder, N., 2011. Peakfinder. Internet: http://www. mathworks. com-/matlabcentral/fileexchange/25500 .

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization, in: Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on, IEEE. pp. 2921–2929.

Zhu, Z., Harowicz, M., Zhang, J., Saha, A., Grimm, L.J., Hwang, E.S., Mazurowski, M.A., 2017. Deep learning analysis of breast mris for prediction of occult invasive disease in ductal carcinoma in situ. arXiv preprint arXiv:1711.10577 .

Zivkovic, Z., 2004. Improved adaptive gaussian mixture model for background subtraction, in: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, IEEE. pp. 28–31.

Zuiderveld, K., 1994. Contrast limited adaptive histogram equalization. Graphics gems , 474–485.

Appendices

Appendix 1

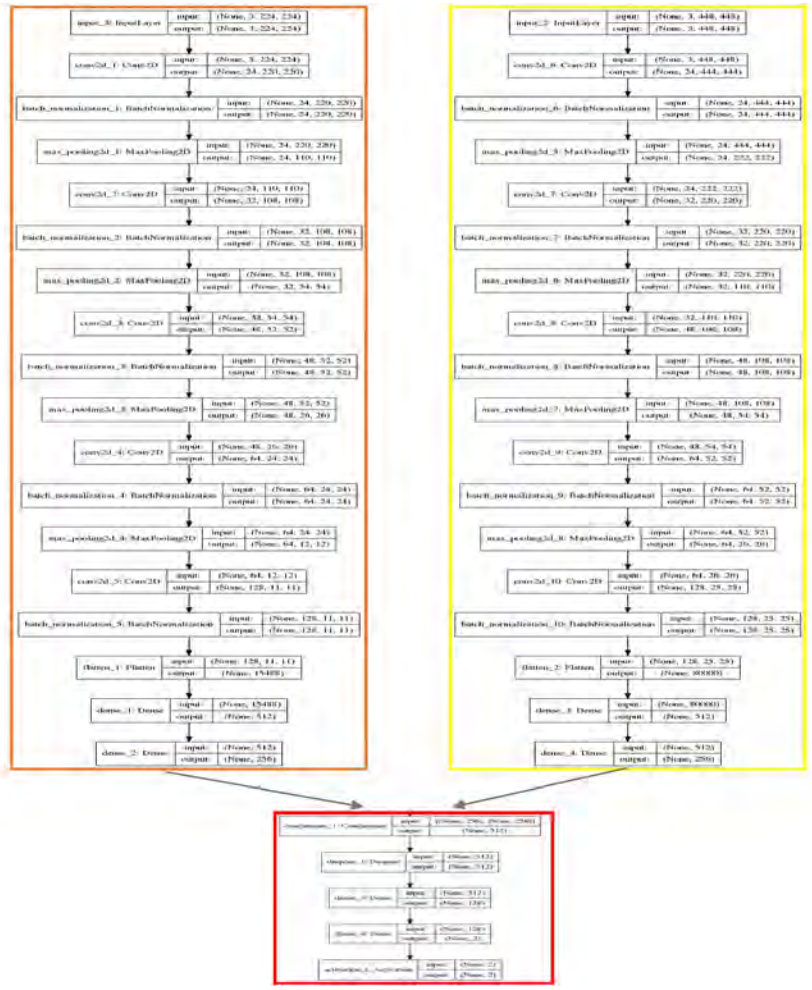


Figure 22: **Bilinear Neural Network (BNN1) Architecture**

Medical Imaging and Applications

Master Thesis, June 2018

Automation of Reflector Delineation for Ultrasound Speed-of-Sound Imaging

Umamaheswaran Raman Kumar
Supervisors: Dr. Sergio Sanabria, Dr. Valery Vishnevskiy, Prof. Dr. Orcun Goksel

Computer-assisted Applications in Medicine (CAiM), ETH Zurich

Abstract

The Computer-assisted Applications in Medicine (CAiM) group from ETH Zurich has proposed a novel reflector-based hand-held Speed-of-Sound (SoS) imaging method for breast cancer screening with minor extensions to conventional ultrasound (US) B-mode systems. The reflector acts as a timing reference for the US signals and its tracking is important for SoS image reconstruction. This thesis work shows an study on automation of reflector delineation for in-vivo US medical images with the previously proposed optimization algorithm based on Dynamic Programming (DP) and provides a comparative study with newly proposed deep learning approach based on simple Convolutional Neural Network (CNN) and U-Net architectures.

Keywords: Ultrasound, Breast cancer, Reflector delineation, Dynamic programming, Deep learning, CNN, U-Net.

1. Introduction

Among U.S. women, breast cancer is the most commonly diagnosed cancer and the second leading cause of cancer death after lung cancer *(Ma and Jemal, 2013)*. It is a high-prevalence disease affecting more women in the U.S and other developed countries. However, over the recent years, breast cancer incidences and mortality have been increasing in developing countries as well. Current routine screening consists of X-ray mammography, which, however, shows low sensitivity to malign tumors in dense breasts, for which a large number of false positives leads to an unnecessary number of breast biopsies. Also, the use of ionizing radiation advises against a frequent utilization, for instance, to monitor the progress of a tumor. Finally, the compression of the breast down to a few centimeters may cause patient discomfort. For these reasons, latest recommendations restrict the general use of X-ray mammography to biennial examinations in women over 50-year-old *(Siu, 2016)*.

Ultrasound is a safe, pain-free, and widely available medical imaging modality, which can complement routine mammographies. The hand-held speed-of-sound imaging method proposed by *Sanabria and Goksel (2016)* transmits US waves through tissue between a B-mode transducer and a hand-held reflector and reconstructs a SoS image of sufficient quality for tumor screening. It only requires a small and localized breast compression, while allowing for flexible access to arbitrary imaging planes within the breast.

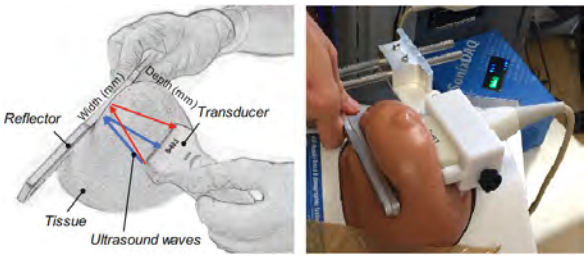
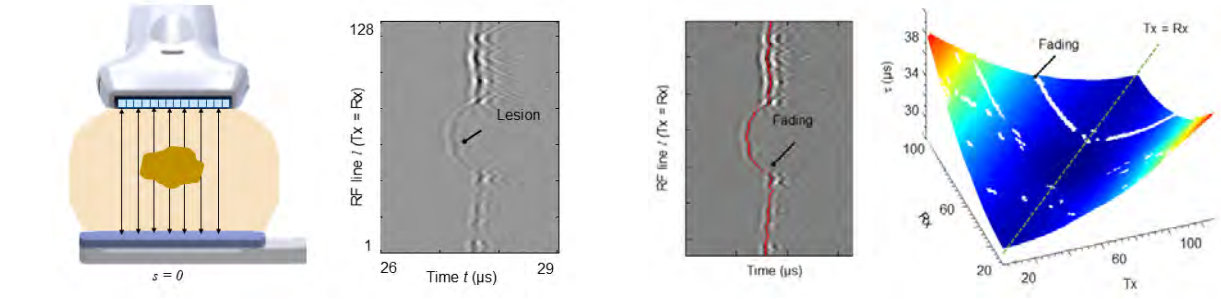


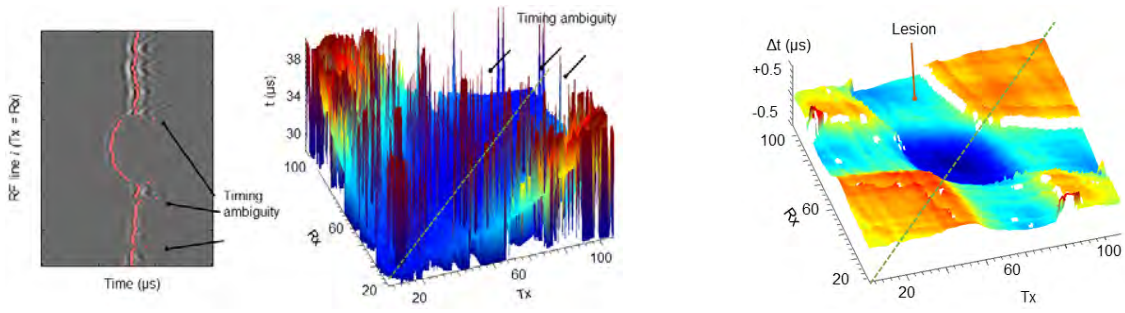
Figure 1: Setup with breast phantom *(Sanabria and Goksel, 2016)*

The SoS image reconstruction uses the hand-held reflector plate as the timing reference, thereby depending highly on the unambiguous measurement of the US time-of-flight (ToF) for a good and reliable reconstruction. The earlier approach uses global optimization based on Dynamic Programming (DP) similar to the one applied for the segmentation of bones *(Foroughi et al., 2007)* and vessel walls in US *(Crimi et al., 2016)*. This



(a) Experimental setup

(c) Global optimization



(b) Individual RF line analysis

(d) Reflector fit

Figure 2: Reflector delineation for ex-vivo liver tissue *(Sanabria and Goksel, 2016)*

approach has been tested on phantoms and ex-vivo samples and has provided promising results. The main work of this thesis was focused to verify the reflector delineation step for in-vivo cases, as the high variability in breast densities affects the tracking, and also to propose a new deep learning approach for reflector delineation to be compared with the existing approach.

2. State of the art

The recorded synthetic aperture dataset $A(t_m, \mathrm{Rx}, \mathrm{Tx})$, with $\mathrm{Rx} = 1 \ldots N$ and $\mathrm{Tx} = 1 \ldots N$, is a 3D matrix consisting of $N \times N$ radio-frequency (RF) lines in function of echo time t_m, where $m = 1 \ldots M$ are discretization indices. A single time-of-flight (ToF) value corresponding to the reflector delineation is searched for each RF line $\boldsymbol{\tau} = \tau(\mathrm{Rx}, \mathrm{Tx})$. Each RF line $A(t_m)$ is a modulated waveform with an oscillatory pattern (Fig. 2a). Due to the heterogeneous SoS distribution, which leads to interference and weak scattering effects, the waveform shape changes for different paths. In the most extreme case, the reflected ultrasound signals may fall below the system noise level (*fading*) at certain paths and a ToF measurement is not possible for the corresponding RF lines.

Therefore, individual RF line analysis, for instance, by picking the peak signal amplitude (Fig. 2b), or by applying more sophisticated correlation-based or wavelet-based methods, inherently lead to timing ambiguities, since different local maxima may be selected for different transmit-receive (Tx-Rx) pairs. On the other hand, waveform demodulation, e.g., by applying a Hilbert transform, leads to a loss of temporal resolution, which distorts the measurement of small perturbations $\boldsymbol{\Delta\tau}$. It is therefore frequent that the calculated ToF matrices in USCT are heavily post-processed to remove timing outliers (Chang et al., 2007; Li et al., 2009; Qu et al., 2015).

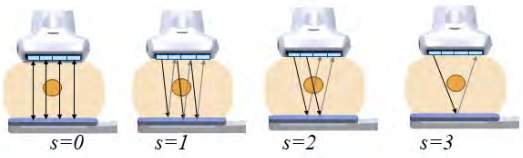


Figure 3: Transmit-receive paths for different shift index s (*Courtesy of Sanabria*)

The proposed global optimization approach, simultaneously evaluates all Tx-Rx traces and minimizes an energy function to calculate the optimum delay matrix $\boldsymbol{\tau}$. This approach reduces timing ambiguities and provides a continuous surface $\boldsymbol{\tau}$ and is used for detecting reflector time delay in RF lines.

Two dimensional algorithm: The algorithm tracks the reflector in a 2D image (B-scan). The horizontal coordinate is the echo time m and the vertical coordinate is a list of successive RF lines l, corresponding to adjacent Tx-Rx pairs. Adjacent pairs show the same lateral Tx-Rx separation $s = \mathrm{Rx} - \mathrm{Tx}$. For instance, Fig. 2a

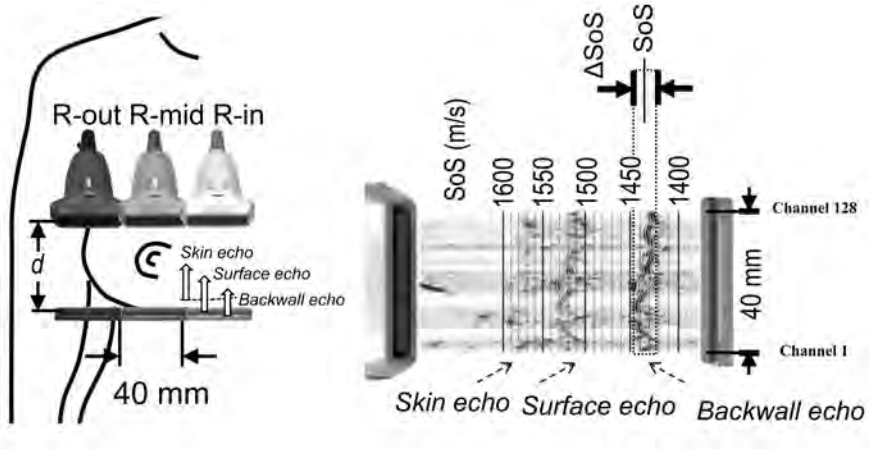


Figure 4: Explanation of multiple echoes observed in in-vivo breast images. Measurements are performed in outer (out), inner (in) and middle (mid) breast segments. The horizontal axis of the US images indicates the echo time of arrival t of the US echoes, while the vertical axis of the US images represents the segment width (probe width or channel index $1 \dots 128$). With a known distance d between probe and reflector, the speed of sound (SoS) is calculated as SoS=0.5*d/t, where the 0.5 factor accounts for the propagation from a transmitter element to the reflector back and forth. Three US echoes can be identified. The first observed echo is a reflection from the breast skin layer. The second echo is generated at the top reflector surface and may consequently overlap in time with the skin surface echo. The third and subsequent echoes correspond to the back wall of the reflector. The median SoS provides a measure of "breast density", where the SoS variation range ΔSoS provides a measure of "breast heterogeneity" ((Sanabria et al., 2018), courtesy of Sanabria)

.

shows the ray paths and a B-scan for $s = 0$ (same element used as both Tx and Rx), and Fig. 3 shows the ray paths for larger s values.For each B-scan $A_{m,l}$, the algorithm cumulatively builds a global cost matrix $C_{m,l}$ along successive RF lines l for each possible timing candidate m. A search window $w = -W/2 \dots W/2$ of W samples is iteratively defined with respect to the adjacent line l. Also, a memory matrix $E_{m,l}$ records discrete timing decisions for each l and m. The optimum reflector delineation minimizes the cumulative cost, and following $M_{m,l}$ backwards the ToF profile $\tau(l) = t_{T(l)}$ is drawn:

$$
\begin{aligned}
C_{m,l} &= \min_{w} \{C_{m+w,l-1} - f_1(A_{m,l}, A_{m+w,l-1})\} - f_0(A_{m,l}) \\
M_{m,l} &= \arg\min_{w} \{C_{m+w,l-1} - f_1(A_{m,l}, A_{m,l-1})\} \\
T(l) &= \begin{cases} \arg\min_{m} C_{m,l} & l = L, \\ M_{T(l+1),l+1} & l = L-1 \dots 1. \end{cases}
\end{aligned} \tag{1}
$$

The cost is evaluated with likelihood f_0 and f_1 smoothness functions in function the current $A_{m,l}$ and adjacent $A_{m+w,l-1}$ B-scan samples. This framework is very general and is used to introduce regularization into the energy function, for instance, in terms of ToF continuity between adjacent Tx index, Rx index pairs, and/or constraints with respect to allowed reflector positions and orientations. In the implementation, f_0 is formulated as a weighted sum of non-linear terms $f_o \propto A_{m,l}, f_{rel}(A_{m,l}), f_{osc}(A_{m,l})$, where f_{rel} and f_{osc} are binary step functions that are respectively activated if A shows a relative maxima at m or if an oscillatory pattern is identified around m. On the other hand, f_1 allows introducing continuity constraints. Also, $f_1 \propto |A_{m,l} - A_{m+w,l}|, w, f_{pj}(A_{m,l} - A_{m+w,l})$, where f_{pj} is a binary step function that is activated if a phase jump $> \pi$ rad occurs between $A_{m,l}$ and $A_{m+w,l}$.

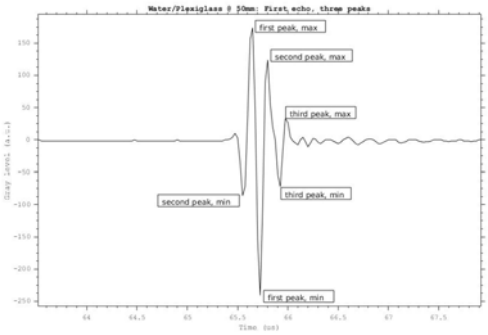
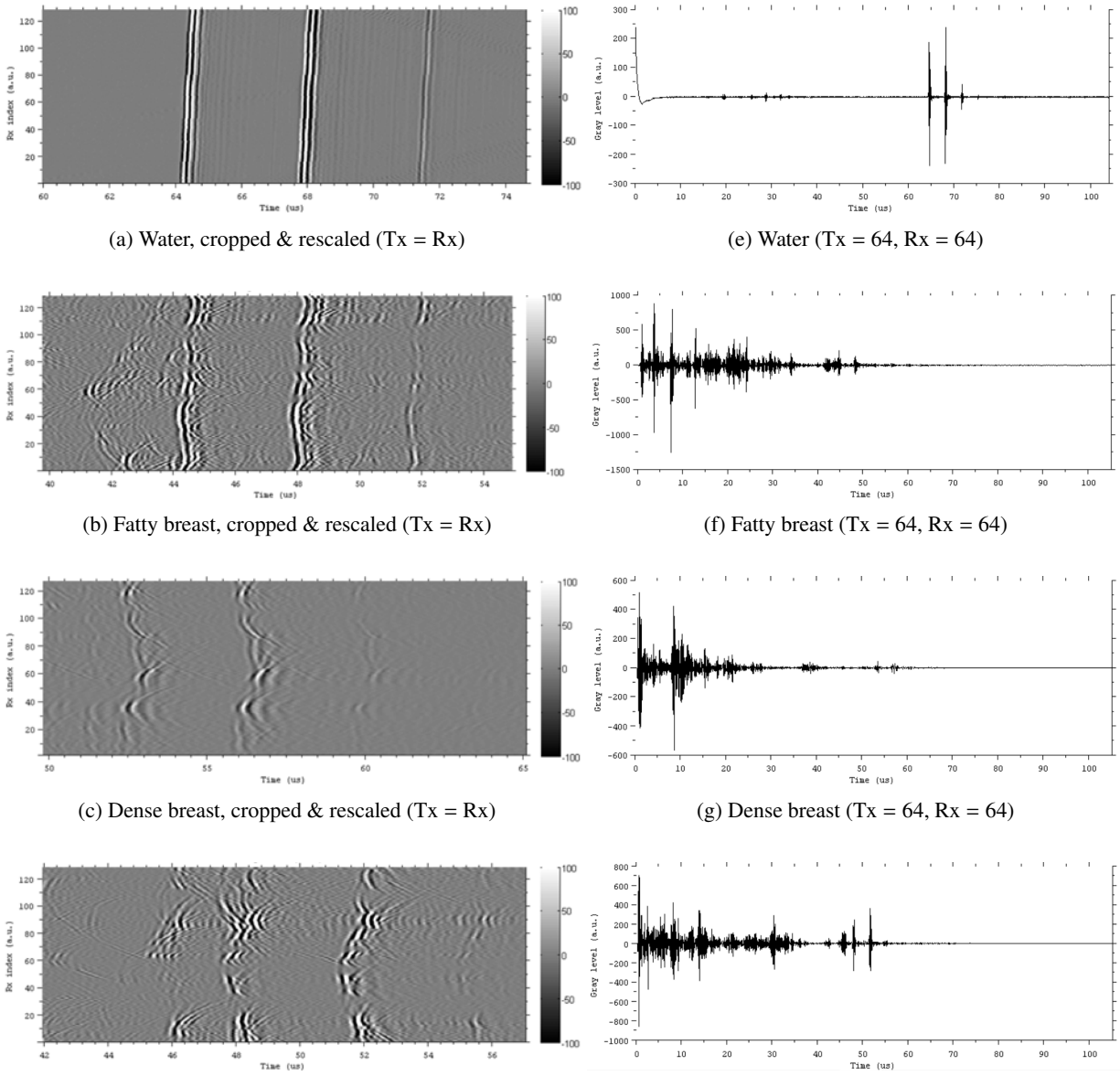


Figure 5: Echo observed for Plexiglas reflector wall in water medium

As observed in figure 5, the ultrasound signal received shows multiple peaks for one echo of the Plexiglas reflector and global optimization using the DP algorithm explained earlier is the current state of the art algorithm used to track the dominating negative peak consistently across all the RF lines. The main drawback of the algorithm is that the equation to be optimized by DP incorporates as regularization information prior information from observed oscillatory patterns from calibration experiments in water. These patterns are representative of phantom and a few ex-vivo cases, where the echoes are easily distinguishable from the rest of the signal, however, they might be difficult to generalize to real heterogeneous breast tissues.



(a) Water, cropped & rescaled (Tx = Rx)

(e) Water (Tx = 64, Rx = 64)

(b) Fatty breast, cropped & rescaled (Tx = Rx)

(f) Fatty breast (Tx = 64, Rx = 64)

(c) Dense breast, cropped & rescaled (Tx = Rx)

(g) Dense breast (Tx = 64, Rx = 64)

(d) Differentially dense breast, cropped & rescaled (Tx = Rx)

(h) Differentially dense breast (Tx = 64, Rx = 64)

Figure 6: Ultrasound data for different breast densities compared to water

As observed in Fig. 6, the recorded ultrasound images show multiple signal packages, corresponding to different time echoes, which are explained in Fig.4. We are here interested in tracking the first reflector echo (surface echo), which gives a clear timing reference for computation of Speed-of-Sound (SoS). However, other reverberations are also present, such as internal tissue reflections at the skin-gland breast layer, and also multiple reflector reverberations. Thus, one main aspect is to be able to discriminate the right oscillation package. This can be partially achieved by cropping a time window of interest on the base of the known distance value d between reflector and probe. However, it is unavoidable that several echo packages are present within this window and the chosen reflector tracking strategy needs also to show discriminatory power.

3. Material and methods

A deep learning approach for segmentation was proposed for reflector delineation to compare and evaluate the performance of the already existing global optimization algorithm using Dynamic Programming (DP). Two networks were implemented as part of this project which aims to segment the reflector echo. The main method is a U-Net based approach, which is a popular biomedical image segmentation technique *(Ronneberger et al., 2015)*, and the second one is a simple Convolutional Neural Network (CNN). The implementation of the network and training was written in Python using Keras API with Tensorflow as back-end, and the data preparation and pre-processing was done using Matlab®.

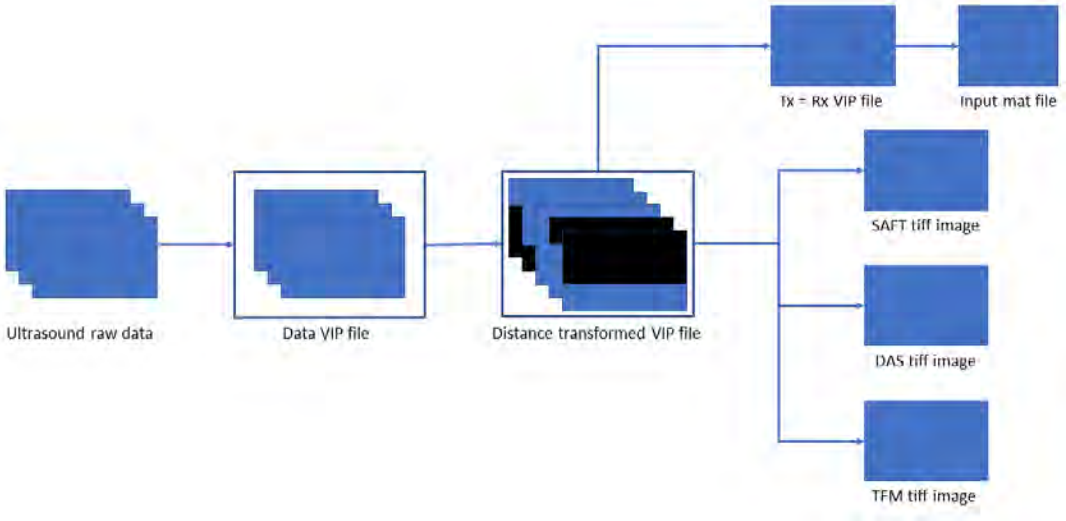


Figure 7: Input data extraction workflow run on cluster

3.1. Data preparation

The breast data was obtained using a 128-element 5 MHz linear ultrasound array (L14/5-38) transducer operated in multistatic mode, each element sequentially transmitting (Tx) an ultrasound pulse and the rest receiving (Rx) the reflected signals. The input data collected from the ultrasound machines are stored as *'.daq'* files and each file contains the data recorded for one line element in the transducer. Every acquisition procedure generates 128 daq files with a total size of $\approx$ 1 GB. Figure 7 shows the workflow for generating different types of processed data from one set of raw data files. The ultrasound data is initially read and stored in one single *mat* file with *'.vip'* extension so that it can be reused easily for further processing. The VIP file contains the data in a 3-dimensional matrix {*time*, *receiver index*, *transmitter index*} with additional header information. The data VIP file is then converted into a distance transformed VIP file {*time*, *transmitter index*, *shift index (distance from transmitter)*}. The shift index s varies from -127 to 127 (Fig. 3) and since all the transmitters do not have that shift, few of the indexes are left with zeros, as shown in Fig. 7 in *'Distance transformed VIP file'* block.

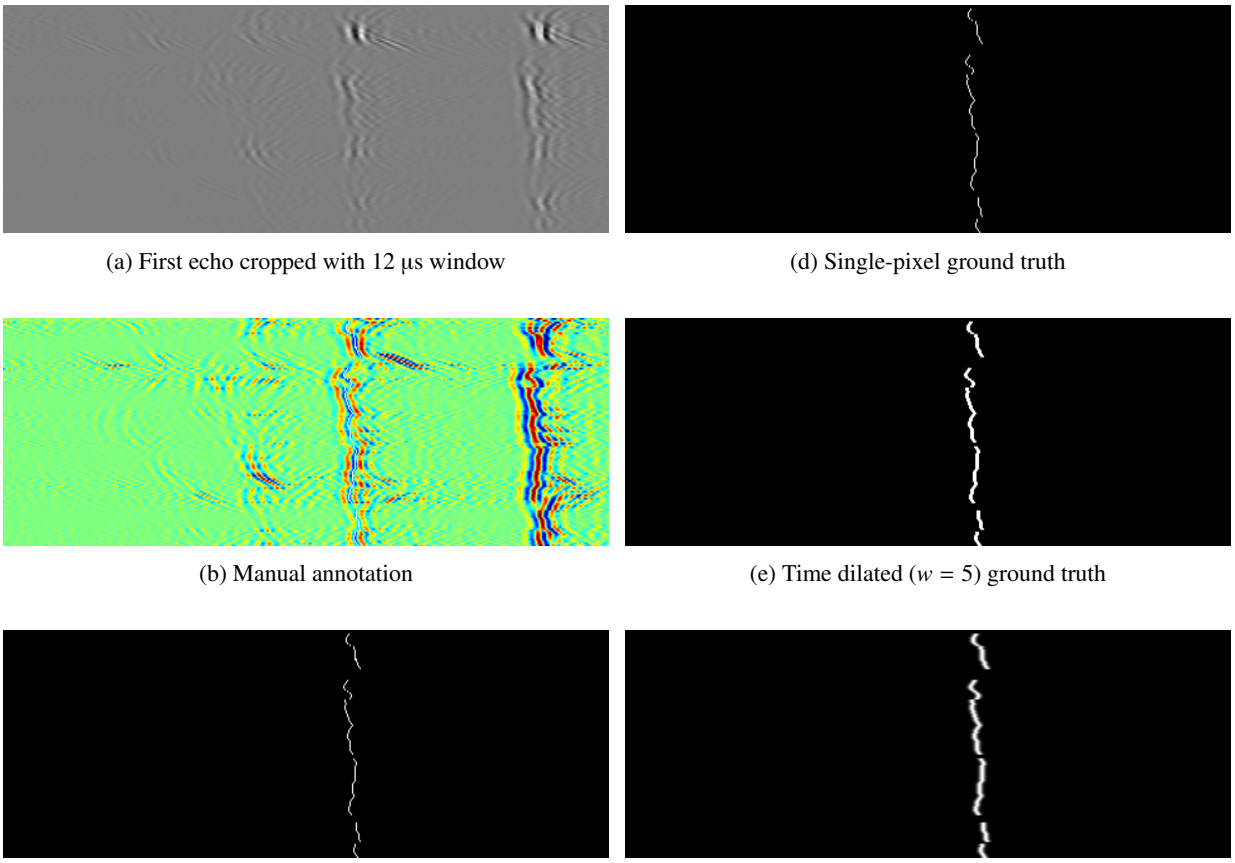
In the experiments, only the s=0 shift data was considered as it gives a complete image of dimension l(128) $\times$ t(4156) with all the 128 lines. Since the distance of the transmitter is already recorded during measurements, it is not required to process the complete image. By using the distance measured between reflector and ultrasound probe and by assuming the nominal speed of sound in soft tissue as 1540 m/s, a rough estimate of the reflector echo time is obtained. Keeping this time as reference, the image is cropped with a window of -6 µs before and 6 µs after the reference time. Since the sampling frequency is $40MHz$, the final image dimension after cropping is 128(height) $\times$ 481(width).

Apart from the input images generated for training the network, several other B-mode images (SAFT, DAS and TFM shown in Fig. 7) were also generated for later experiments. Previously, the entire workflow was run manually using a GUI application which had the limitation to process only one set of acquisition data at a time and also it took $\approx$ 25 mins to process one set of files. Since there were 950 set of files to process, it would take more than 2 weeks to generate all the required data. Therefore, few parts of the Matlab code were optimized and then converted to *'mex'* files to run faster. This took $\approx$ 9 mins from start to end in order to process the same set of files on a SGE (Oracle Grid Engine) cluster.

3.2. Ground truth generation

The images generated in section 3.1 were annotated by 2 readers ($R1$ by clinician and $R2$ by -the author of this work) to get 2 sets of annotated data that can later measure the performance of the network trained by different annotators. Annotations were done on the jet scale images (Fig. 8a) which are more easier for visual interpretation compared to gray scale images for manual annotators. Fig. 8b and 8c shows the manual annotated image and its corresponding annotation extracted. A total of 398 images collected for a previously published breast density classification study (Sanabria et al., 2018) were annotated by both readers and another 124 (out of 450) images collected for menstrual cycle study were annotated only by the clinician ($R1$).

Three different ground truth images are generated from the annotated images for training the network and



(a) First echo cropped with 12 µs window

(d) Single-pixel ground truth

(b) Manual annotation

(e) Time dilated ($w = 5$) ground truth

(c) Manual annotation extracted

(f) Time dilated gaussian ($w = 9$) ground truth

Figure 8: Ground truth generation

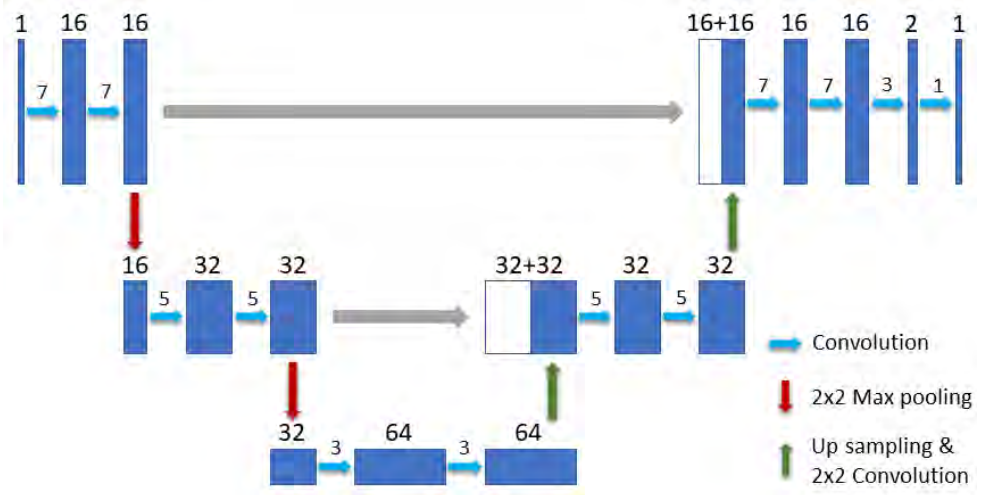
to understand its behaviour. The first ground truth image (single-pixel) (Fig. 8d) is created by keeping only one pixel for every line element annotated by selecting the left most pixel from the annotated pixels. The other two ground truth images(Fig. 8e and 8f) were created from the previous ground truth images by applying a box filter (window size, $w = 5$) and a gaussian filter (window size, $w = 9$ and standard deviation, $\sigma = (w - 1)/5 = 1.6$) respectively along the line. The filter sizes were dimensioned to cover half an oscillation period of the signal (4 pixels), which represents the manual annotation accuracy. The box filter gives a dilated image with equal weighting to the dominating echo oscillation pixels, whereas the gaussian filter gives a smooth dilation with more weighting to the center of the dominating peak. It should be noted that once the right oscillation period has been identified with any method, sub-pixel accuracy can be obtained by applying state-of-the-art interpolation methods (for instance polynomial fitting (Azar et al., 2010).

3.3. Network architecture

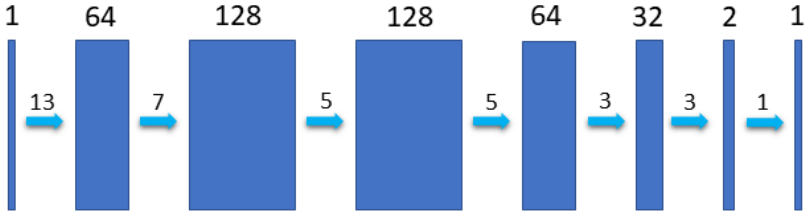
Two different network architectures are proposed with the goal to segment only the first echo of the reflector as close as possible to the ground truth images. The decision to use CNN's instead of any simple classifier such as SVM or k-Nearest Neighbour is because the reflector echo shows large variance in structure and characteristics for different breast densities and it is not easy to model a classifier for this dataset with statistical or hand-picked features.

3.3.1. U-Net architecture

The network architecture mainly used in this work is a U-Net based architecture originally proposed by *Ronneberger et al. (2015)* for biomedical image segmentation. The major advantage provided by using this architecture is that it is known to work well with small training datasets. Minor design changes were made to the original U-Net implementation in order to work with a dataset of less than 500 images. Fig. 9a shows the final chosen implementation after trying a few combinations of downsampling layers with different filter sizes and number of feature maps for each layer. The input layer dimension (128×484) is chosen to be a divisible of four as there are two max-pooling layers, so the input images were zero-padded to get this dimension. All the convolution layers uses *ReLU* activation functions except the last layer which uses *sigmoid* activation function to generate an output probability map. Larger convolution



(a) U-Net architecture



(b) Simple CNN architecture

Figure 9: Network architectures

filter sizes were used for larger feature maps in order to capture more information from the neighboring pixels.

3.3.2. *Simplified CNN architecture*

Fig. 9b shows the network architecture of a simplified CNN structure which was used for comparison to U-net. Contrary to the U-Net architecture, there are no down-sampling or up-sampling layers in this architecture. Thus all the layers are 2*D-convolution layers* with *ReLU* activation functions except the last layer which uses a *sigmoid* activation function. The top layers in the network use larger filter sizes and the size of the filter is gradually decreased as it goes deeper into the network. The size of each feature map is kept constant at every layer, therefore, there is no constraint on the input layer size and hence the input images are not padded.

3.4. *Network Training*

The networks were trained with the breast density study dataset consisting of a total of 398 images. The breast density dataset categorizes the breasts into four categories based on the percentage of glandular (mammary) tissue. The reference standard for classification was obtained by additionally acquiring X-ray mammographies for each dataset, which were evaluated according to the American College of Radiology (ACR) Breast Imaging Reporting and Data System (BIRADS) 5th edition. This classification system, which is used widely, defined four qualitative breast-density categories: almost entirely fatty (A), scattered areas of fibroglandular density (B), heterogeneously dense (C) and extremely dense (D) (Sanabria et al., 2018). The dataset was acquired for women of a broad age range, with an average of 56.5 years and a range of 34–85 years. Different densities of breasts show variable characteristics for reflector echoes, and so, each category of the dataset was split into 90% training data and 10% validation data as shown in Table 1.

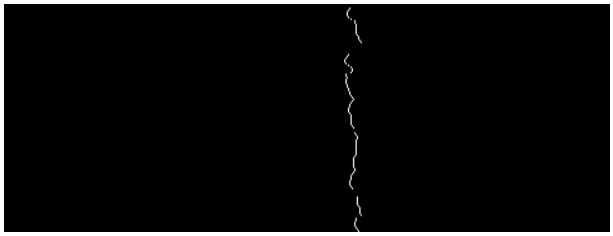
Table 1: Data split for training and validation

Breast Category (Density)	**Total**	**Training** 90%	**Validation** 10%
A (0 – 25%)	73	66	7
B (25 – 50%)	122	110	12
C (50 – 75%)	66	59	7
D (75 – 100%)	28	25	3
n.s. (not specified)	109	98	11
	398	358	40

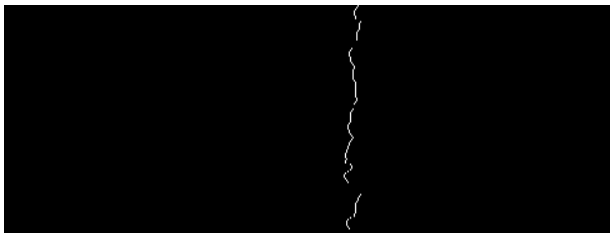
A second dataset, named "menstrual cycle" was more recently measured and annotated by only one reader R1. This dataset contains 120 images of young women (mean age 26.6 years with a range of 18–40 years) and predominantly dense breast categories (65 D, 30 C, 6 B and 19 A). The dataset was used to further validate the trained model for unseen data. Since both the datasets had different set of patients for conducting the study, it also helps to ensure the network is not over-fitted.

3.4.1. Data augmentation

From a CNN perspective, the breast density dataset is a relatively small dataset with only 398 annotated images. In order to increase the amount of training data, several homogeneous transformations can be applied on the training images and labels. It is important to understand the characteristics of the data before performing any augmentation, as all the augmentations may not introduce artifacts in the network. In this case, vertically flipping the image is a valid augmentation that can be applied to the dataset as shown in fig. 10 because it preserves the oscillation pattern of the echo in every line and also the continuity between adjacent channel lines.



(a) Original ground truth image



(b) Vertically flipped ground truth image

Figure 10: Data augmentation

3.4.2. Optimizer

The Adam optimizer, as proposed by *Kingma and Ba (2014)*, is a common variant to stochastic gradient descent optimizer and it was here used for all the trainings and both networks. The method computes individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients. The initial learning rate of 10^{-3} was found to perform better compared to other learning rates in the range between $10^{-1} - 10^{-4}$.

3.4.3. Loss function

There are several different loss functions available and the selection of the right loss function used for optimization is necessary to ensure the convergence of the network during the training. The different loss functions used for training the network are provided with the actual equations used for implementation. $y_{true,n}$ is the ground truth label for individual pixels and $y_{pred,n}$ is the label predicted by the network and N is the total number of pixels of all the images in a mini-batch.

i *Mean squared error (MSE)*

$$MSE = \frac{1}{N}\sum_{n=1}^{N}(y_{true,n} - y_{pred,n})^2 \tag{2}$$

ii *Dice coefficient error (DCE) (Milletari et al., 2016)*

$$DCE = 1 - \frac{2 \times \sum_{n=1}^{N} |y_{true,n} \times y_{pred,n}|}{\sum_{n=1}^{N} y_{true,n}^2 + \sum_{n=1}^{N} y_{pred,n}^2} \tag{3}$$

iii *Jaccard coefficient error (JCE)*

$$JCE = 1 - \frac{\sum_{n=1}^{N} |y_{true,n} * y_{pred,n}|}{\sum_{n=1}^{N} (|y_{true,n}| + |y_{pred,n}| - |y_{true,n} * y_{pred,n}|)} \tag{4}$$

iv *Custom binary cross entropy (CBCE)*

$$CBCE = -\sum_{n=1}^{N} \{\alpha y_{true,n} log(y_{pred,n}) + \beta(1 - y_{true,n}) log(1 - y_{pred,n})\} \tag{5}$$

where, α and β are the misclassification costs for class '1' label and class '0' label respectively.

Among all the above loss functions tested, the $CBCE$ loss function (Eq. 5) was the only one that reached convergence. The main reason is that the number of class '0' labels were far greater than class '1' labels and the $CBCE$ loss function was able to account for this imbalance by weighting the misclassification cost of the two classes differently. The misclassification of class '1' was given more weightage compared to class '0'. The final values of α and β used in eq. 5 are 25 and 1 respectively. The value of α could be potentially increased until approx. 50. However, as it is shown afterwards, this would then classify more pixels as class '1', leading to more false positives in spurious echo signals (for instance, breast skin echoes or multiple reflector reverberations), and increasing the need of post-processing to remove these. Therefore, α is chosen as a trade off between sensitivity of reflector detection and specificity to spurious echo signals.

3.4.4. Callback function

Three different keras callback functions were included during the network training. The *Tensorboard* callback was used for debugging the network during the process of training, the *CSVlogger* callback was used to store the values of the loss functions and metrics for each epoch and the *Model checkpoint* callback was used to store the weights of the network after every epoch and only if there is a decrease in the loss function.

3.5. Post processing

The output of the networks are probability maps and in order to get a binary segmentation mask some post-processing is applied. The binarization of the output could also be implemented as the last layer of the network but it is presently not utilizing in order to test different post-processing strategies. The first step is to remove the probability values lower than 0.1 because the output of the network has many non-zero values in this range close to the echo. In the next step, one maximum value is selected for every line l and if more than one value is found then the median of those values is selected. After finding the maximum values for every line, the median over all lines is calculated which is the considered as the median index of the images and this connects to the average SoS measurement for each segment as plotted in Fig. 4. Finally, a box filter is applied to remove all outliers outside a window of 3µs (120 indexes) centering the median index value calculated. Then the echo time value for each line is re-calculated as the median of the maximum probability values in the cropped search window.

3.6. Evaluation Metrics

Several metrics were used for evaluating the training of the networks and also for comparing the performance of the current method implemented with the previous DP approach. The various metrics used for evaluation are explained with the actual equations used for implementation.

3.6.1. Binary accuracy (%)

The equation for binary accuracy is given by,

$$Accuracy_{binary} = \frac{1}{N}\sum_{n=1}^{N}|y_{true,n} - y_{pred,n}| \times 100 \qquad (6)$$

The binary accuracy is the main metric used for evaluating the network training and it is evaluated both for training and validation data to observe if the network is over-fitting. The metric is calculated after binarizing the output probability map from the network with a threshold value of 0.5. For training, it is calculated for every mini-batch data and averaged over all the batches of an epoch and for validation it is calculated over the validation data after every epoch.

3.6.2. Dice similarity coefficient (DSC)

The equation for the dice coefficient is given by,

$$DSC = \frac{2 * |I_{true} \cap I_{pred}|}{|I_{true}| + |I_{pred}|} \qquad (7)$$

The Dice Coefficient is calculated only for class '1' to get a better insight of the segmentation as the overall dice coefficient is very biased because of the domination of class '0'. Both the ground truth and predicted masks are dilated for 5 pixels in the horizontal direction before calculating the dice in order to capture the complete dominant peak oscillation, and find the overlap without the manual annotation uncertainty region.

3.6.3. Time difference (µs)

The main aim of the thesis work is to track the reflector echo time for Speed-Of-Sound computation purposes. Therefore, it is important to calculate the difference in time between the predicted and ground truth values. The measure of the difference of time is calculated both pixel-wise, in order to asses outliers, and image-wise(median of pixel-wise difference per image), in order to assess the time uncertainty when calculating the average SoS. These error metrics are also calculated with both absolute difference and signed difference to know if the prediction of the first echo is biased more by the skin echo or the second reflector echo. Since, the sampling frequency of the signal is $40MHz$ and to get the metric result in time units (µs), the difference calculated pixel-wise or image-wise in indexes is divided by 40.

3.6.4. Mean absolute error (MAE) (µs)

The MAE is calculated on the image-wise time difference for N images and it is given by,

$$MAE = \frac{1}{N}\sum_{n=1}^{N}|\Delta t_{pred,n} - \Delta t_{true,n}| \qquad (8)$$

3.6.5. Root-mean-square error (RMSE) (µs)

The RMSE is calculated on the image-wise time difference for N images and it is given by,

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\Delta t_{pred,n} - \Delta t_{true,n})^2} \qquad (9)$$

3.6.6. Execution time (s)

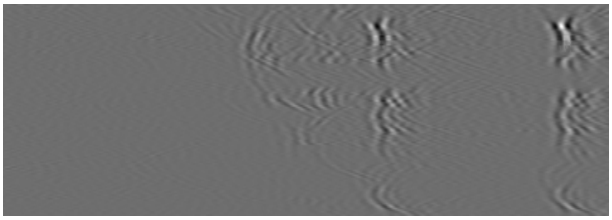
Execution time metric is used to compare the time taken by the network to predict one image, with the time taken by the DP approach for the same image. This helps to identify the approach which is more suitable for real time application.

4. Results

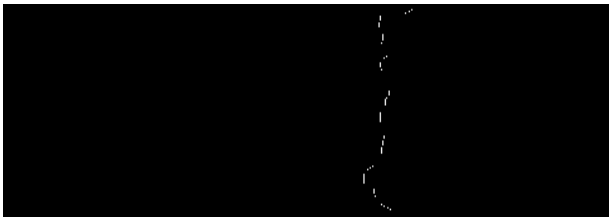
This section provides the results obtained from experiments conducted using U-Net, CNN and DP with optimized parameters and compares their performance.

4.1. Qualitative results

Input and output images are shown for datasets corresponding to different breast densities such as fatty (fig. 12), differentially dense (fig. 13) and dense (fig. 14 & 15). Output images are calculated with both global optimization using DP approach (c), and with U-Net (e) trained with the single-pixel ground truth images (fig. 8d) based on $R1$ annotations. The results from other types of ground truth images (time dilated , fig. 8e, and time dilated Gausssian, fig. 8f) are not provided, as they were not comparable with the rest of the methods.



(a) Original image1

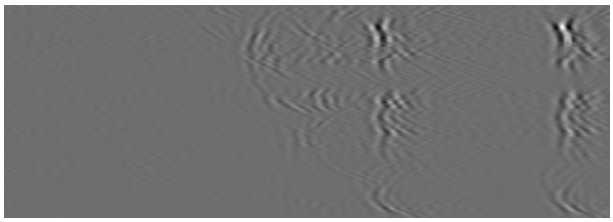


(b) Ground truth image

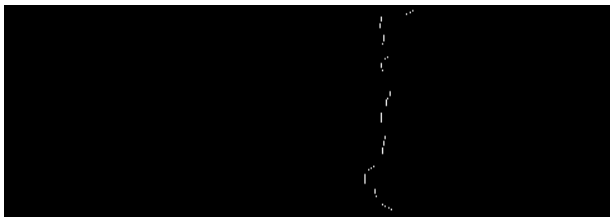
(c) Output image from U-Net

(d) Output image from CNN

Figure 11: Comparison of ground truth and predicted images (U-Net, CNN) for fatty breast (same example as in Fig. 12)

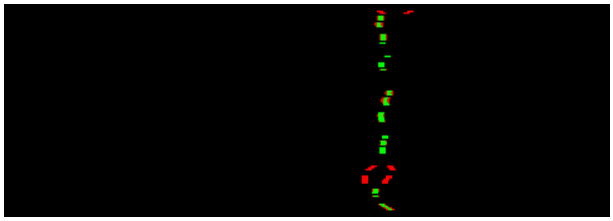


(a) Original image



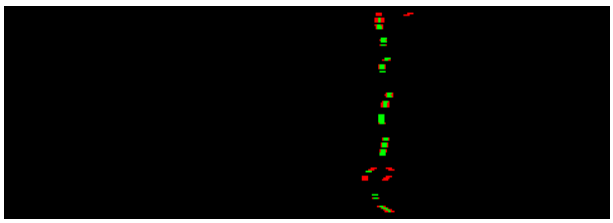
(b) Ground truth image

(c) Output image from DP



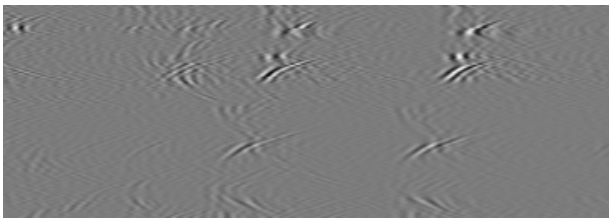
(d) Overlapped image - GT & DP (Green - overlap)

(e) Output image from U-Net

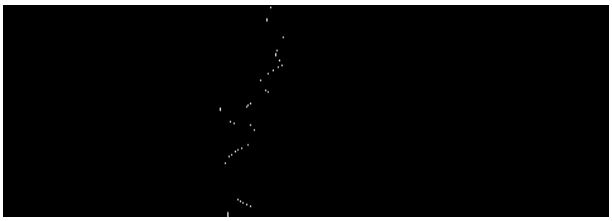


(f) Overlapped image - GT & U-Net (Green - overlap)

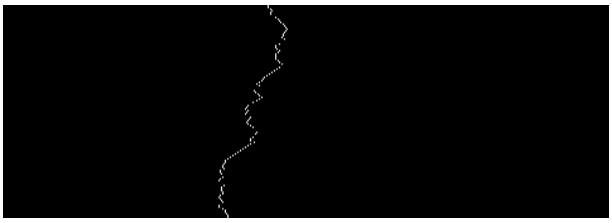
Figure 12: Comparison of ground truth and predicted images (DP, U-Net) for fatty breast



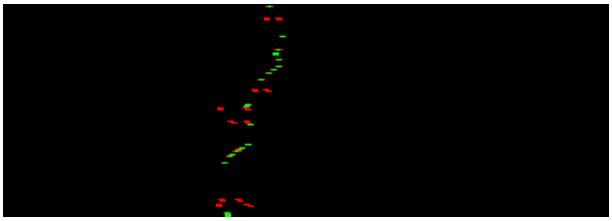
(a) Original image



(b) Ground truth image

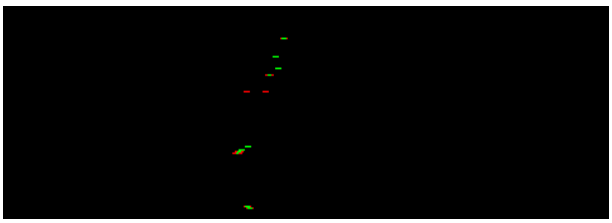


(c) Output image from DP



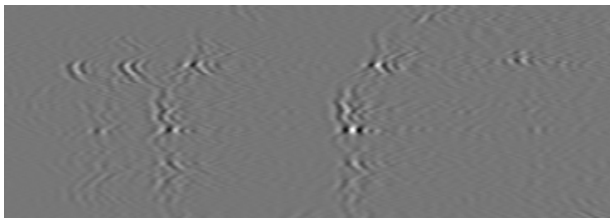
(d) Overlapped image - GT & DP (Green - overlap)

(e) Output image from U-Net



(f) Overlapped image - GT & U-Net (Green - overlap)

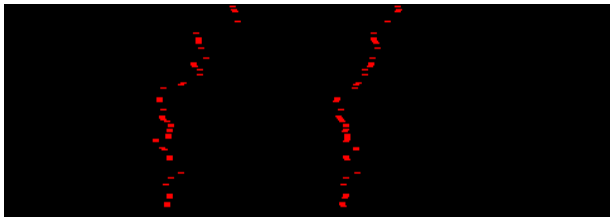
Figure 13: Comparison of ground truth an predicted images (DP, U-Net) for differentially dense breast



(a) Original image

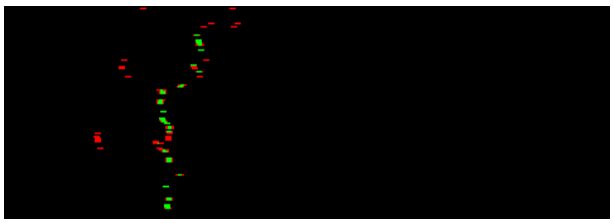
(b) Ground truth image

(c) Output image from DP



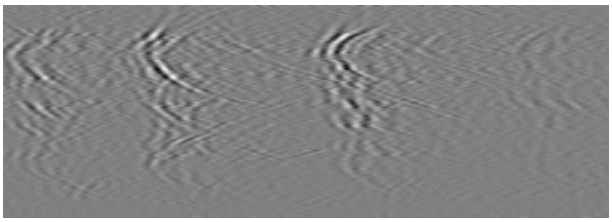
(d) Overlapped image - GT & DP (Green - overlap)

(e) Output image from U-Net



(f) Overlapped image - GT & U-Net (Green - overlap)

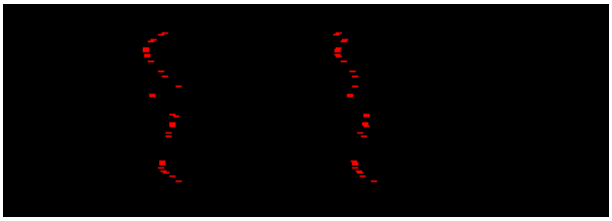
Figure 14: Comparison of ground truth an predicted images (DP, U-Net) for dense breast. In this case, DP wrongly selects the second reflector echo, so that there is no overlap with the ground truth. U-Net correctly selects the first echo



(a) Original image1

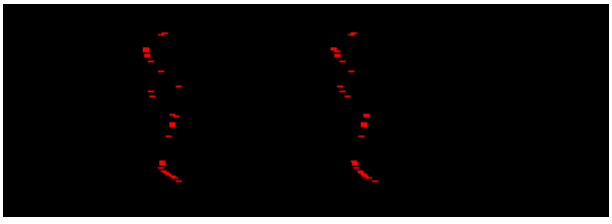
(b) Ground truth image

(c) Output image from DP



(d) Overlapped image - GT & DP (Green - overlap)

(e) Output image from U-Net



(f) Overlapped image - GT & U-Net (Green - overlap)

Figure 15: Comparison of ground truth an predicted images (DP, U-Net) for dense breast (2^{nd} example). In this case, both DP and U-Net wrongly select the second reflector echo, so that there is no overlap with the ground truth.

4.2. Quantitative results

The quantitative results section provides all metrics defined in section 3.6. Fig. 16 & 17 shows the loss and accuracy graphs for both training and validation of the U-Net trained with reader $R1$ ground truth images. Fig. 18-24 show box plots of the percentile distribution of the Dice and distance metrics for pixel-wise and image-wise comparison between different methods. The consensus between the two readers is also plotted in all figures as a reference. For these plots, the U-Net was trained either with the ground truths of reader $R1$ or reader $R2$. For metrics evaluation, the predicted images were compared to the ground truth images of both readers. For example, 'R1-UNET R2' compares the validation ground truth images annotated by reader 1 with the predictions of U-Net trained using the ground truth images of reader 2.

4.2.1. Training graphs

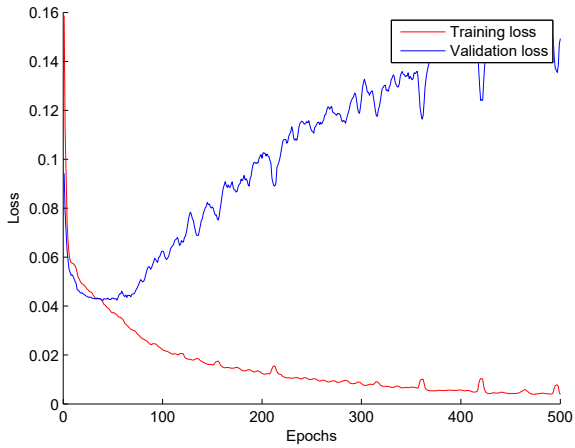


Figure 16: Training and validation loss for U-Net

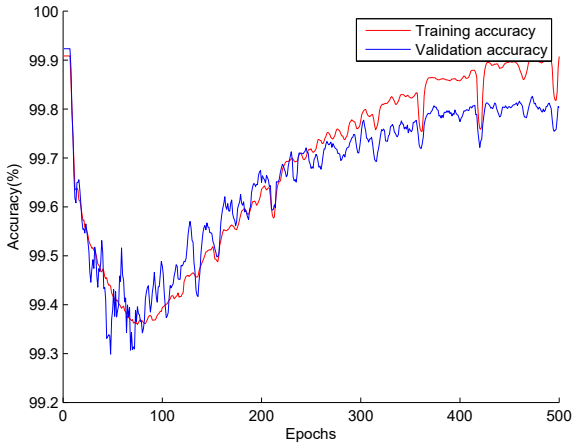


Figure 17: Training and validation accuracy for U-Net

4.2.2. *Dice coefficient*

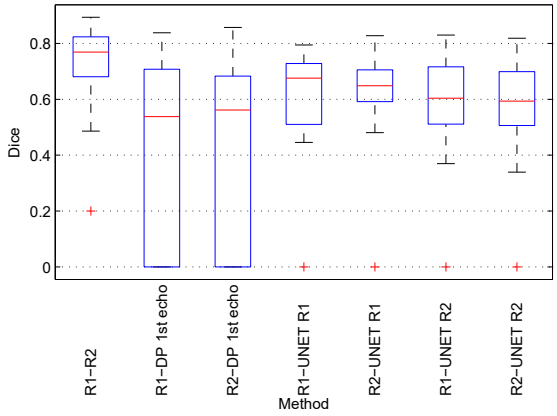


Figure 18: Dice coefficient

4.2.3. *Pixel-wise time difference*

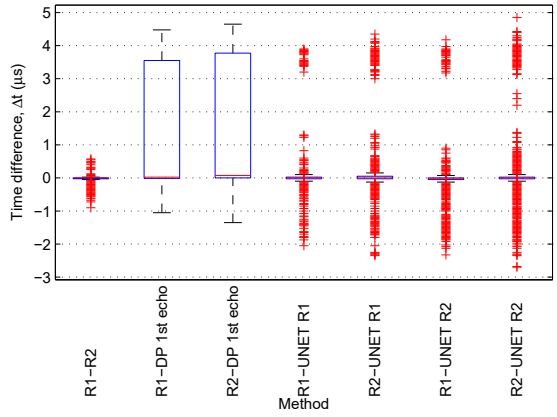


Figure 19: Pixel-wise time difference

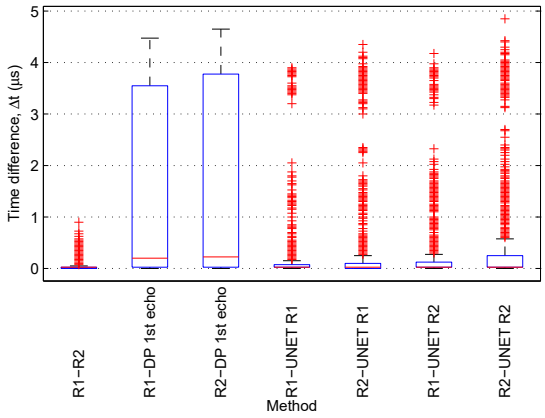


Figure 20: Pixel-wise time difference (absolute)

4.2.4. *Image-wise time difference*

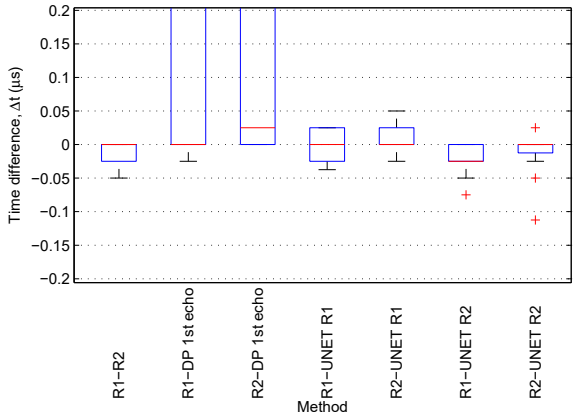


Figure 21: Image-wise time difference

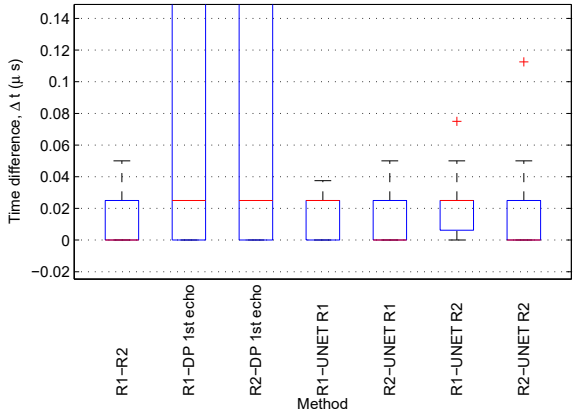


Figure 22: Image-wise time difference (absolute)

4.2.5. *Mean absolute error (MAE)*

Table 2: MAE for different methods

Method	**MAE (μs)**
R1 - R2	0.0181
R1 - DP (1^{st} echo)	1.4881
R2 - DP (1^{st} echo)	1.5003
R1 - UNET R1	0.5528
R2 - UNET R1	0.4709
R1 - UNET R2	0.3959
R2 - UNET R2	0.3841

4.2.6. Root-mean-square error (RMSE)

Table 3: RMSE for different methods

Method	RMSE (μs)
R1 - R2	0.0576
R1 - DP (1^{st} echo)	2.3454
R2 - DP (1^{st} echo)	2.3656
R1 - UNET R1	1.3979
R2 - UNET R1	1.3100
R1 - UNET R2	1.1515
R2 - UNET R2	1.1828

4.3. Performance of U-Net on unseen menstrual dataset

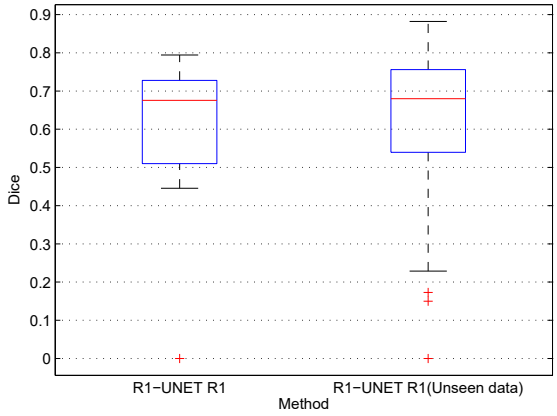


Figure 23: Dice coefficient

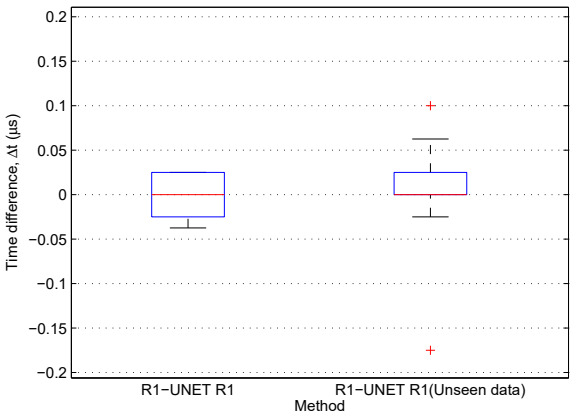


Figure 24: Image-wise time difference

Also, the MAE and RMSE values are 0.3838 and 1.1138 respectively for U-Net prediction for unseen menstrual data.

4.4. Training & Execution time

The time taken by the U-Net described in section 3.3.1 takes ≈ 6 hours for training with 398 images for 500 epochs on a *Nvidia Titan Xp 12GB GPU* machine and it takes ≈ 0.025 s to predict one image, whereas the time taken by the DP algorithm is to predict one image is ≈ 1 s on a *Dual core 16GB RAM CPU* machine.

5. Discussion

From the qualitative images (section 4.1), it can be observed that both methods, DP and U-Net, are able to localize the first echo peak significantly better than CNN. DP always localizes the dominant peak (either first or second) to just 1 pixel. On the other hand, U-Net localizes the first peak and sometimes the second peak to 1-3 pixels and CNN localizes to 2-5 pixels without any post-processing.

In the case of breast density data, both DP and U-Net were able to track the first echoes for fatty breasts (Fig. 12) with more dominant first echo. However, when the second echo becomes more dominant, the DP always tracks the dominant echo and not the first echo. On the contrary, U-Net has higher probability of tracking the first echo for few cases even with other dominant echoes (Fig. 13 and 14). There are also cases (Fig. 15) where both the methods fail to track the first echo. It is also observed that the echoes tracked by DP is more smooth and has less inconsistency along the line elements when compared to the echoes tracked by U-Net.

Fig 16 and 17 shows the graphs for the training loss and accuracy of U-Net for 500 epochs. It can be observed that the loss function is not entirely representative of the problem because it looks like the network is over-fitted with the training data whereas the accuracy graph on the validation set seems to increase. It is due to the fact that the loss function used is an entropy function that is based on probability. When the network tries to assign few pixels closer to 1, it decreases the probability of class 0 pixels. These small changes in probability has a large impact on the final loss because there are many pixels belonging to class 0. On the other hand, the metric used for evaluation cannot account for this as it is a binary accuracy function that binarizes the output with a threshold of 0.5 before calculating the accuracy. Fig 23 and 24 shows that the network does not over-fit as the dice and time difference metrics works good even for unseen 'menstrual study' dataset. Additionally, the MAE and RMSE metrics for the network validated with 'menstrual study' dataset are approximately equal to the 'breast density study' dataset.

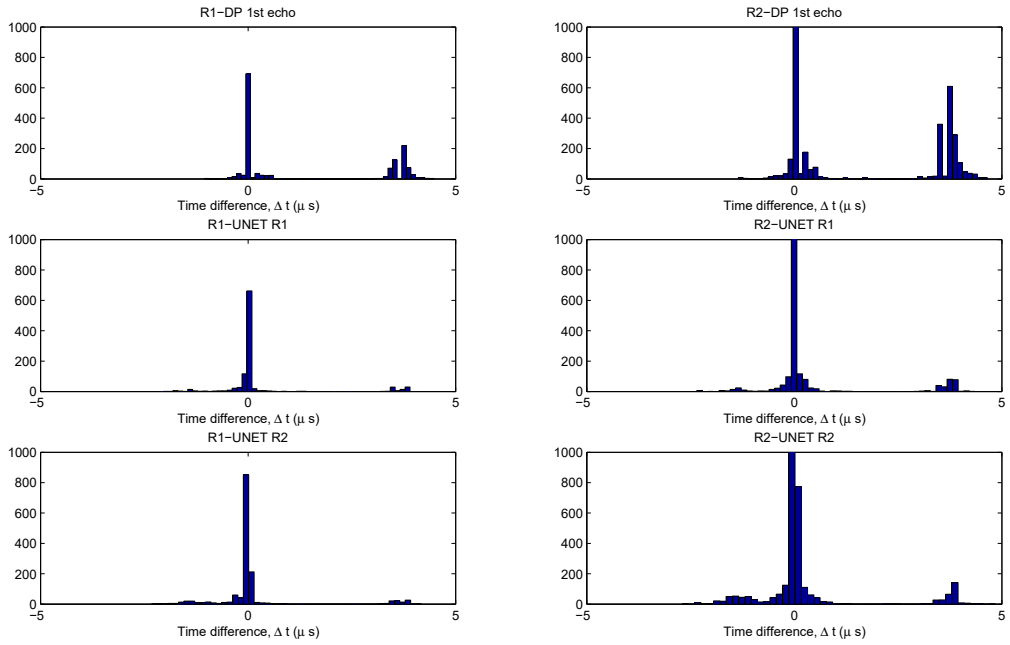


Figure 25: Histogram distribution for different methods for pixel-wise time difference (50 bins)

From the box plots (18-22), it is observed that there is a certain level of disagreement even between the two readers which gives a top boundary of achievable performance. When comparing the medians in all the box plots, the U-Net method trained with the ground truth images from reader 1 is relatively close to the manual annotator compared to the rest of the methods. It can also be observed that DP fails for many cases due to frequent false hits in second echo than U-Net. This can be seen more clearly from Fig. 25 which shows the histograms for the image-wise time difference metric, wherein the second histogram peak with a distance of $\approx$ +4 µs is significantly high for DP compared to the other methods.

6. Conclusions

This thesis proposed a deep learning approach based on simple CNN and U-Net architectures for the automation of reflector delineation for in-vivo US medical images and compared it against the previous state-of-the-art global optimization algorithm based on DP.

Based on the study, the DP approach works well in tracking the dominant echo but it fails to track the first echo in the presence of a dominant second echo. The newly proposed U-Net based approach works relatively better than the DP based approach. Additionally, the execution time of the U-Net based approach outperformed the DP based approach (by 40x) for real time application unless significant implementation changes are made to the DP algorithm to run on a GPU machine.

The future works can involve: (1) Using pre-processed images (making the first echo dominant) for training the U-Net. When the input images were pre-processed, DP showed remarkable improvement in tracking the right echo, almost equal to manual annotators. The pre-processing involves overlapping the two reflector echoes according to their known separation (pre-defined based on the thickness of the reflector), thereby making the first echo more dominant. This is expected to improve the performance by suppressing the skin echo and second echo. (2) Training the network with the whole $3D$-data ($s = 1 \ldots N$), instead of only using the $2D$ images ($s = 0$) can enhance the performance. (3) Using the output probability map of U-Net as input weights for the optimization using DP. U-Net is comparatively robust in tracking dominant echo and DP tracks the echo more continuously along all the line elements, it might be a very good approach going forward.

7. Acknowledgments

Firstly, I would like to thank all my supervisors Dr. Sergio Sanabria, Dr. Valery Vishnevskiy and Prof. Dr. Orcun Goksel for their support and guidance throughout this thesis work. I would also like to acknowledge the support of Dr. med. Ruby from the Institute of Diagnostic and Interventional Radiology at the University Hospital of Zurich for the annotation of the clinical breast

data. I would also like to thank other members of the Computer Vision Lab (CVL), ETH Zurich, who were involved in sharing their knowledge on the work carried out in regard to this thesis. I would like to acknowledge European Commission for funding my studies. Finally, I must express my very profound gratitude to my family and all my friends for providing me with unfailing support and continuous encouragement throughout my years of study.

References

Azar, R.Z., Goksel, O., Salcudean, S.E., 2010. Sub-sample displacement estimation from digitized ultrasound RF signals using multi-dimensional polynomial fitting of the cross-correlation function. IEEE Trans Ultrason Ferroelectr Freq Control 57, 2403–2420.

Chang, C.H., Huang, S.W., Yang, H.C., Chou, Y.H., Li, P.C., 2007. Reconstruction of ultrasonic sound velocity and attenuation coefficient using linear arrays: Clinical assessment. Ultrasound in Medicine and Biology 33, 1681–1687.

Crimi, A., Makhinya, M., Baumann, U., Thalhammer, C., Szekely, G., Goksel, O., 2016. Automatic measurement of venous pressure using b-mode ultrasound. IEEE Transactions on Biomedical Engineering 63, 288–299.

Foroughi, P., Boctor, E., Swartz, M.J., Taylor, R.H., Fichtinger, G., 2007. P6D-2 ultrasound bone segmentation using dynamic programming, in: Ultrasonics Symposium, 2007. IEEE, IEEE. pp. 2523–2526.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. CoRR abs/1412.6980. URL: http://arxiv.org/abs/1412.6980, arXiv:1412.6980.

Li, C., Huang, L., Duric, N., Zhang, H., Rowe, C., 2009. An improved automatic time-of-flight picker for medical ultrasound tomography. Ultrasonics 49, 61–72.

Ma, J., Jemal, A., 2013. Breast Cancer Statistics. Springer New York, New York, NY. pp. 1–18. URL: https://doi.org/10.1007/978-1-4614-5647-6_1, doi:10.1007/978-1-4614-5647-6_1.

Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 3D Vision (3DV), 2016 Fourth International Conference on, IEEE. pp. 565–571.

Qu, X., Azuma, T., Imoto, H., Raufy, R., Lin, H., Nakamura, H., Tamano, S., Takagi, S., Umemura, S.I., Sakuma, I., et al., 2015. Novel automatic first-arrival picking method for ultrasound sound-speed tomography. Japanese Journal of Applied Physics 54, 07HF10.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

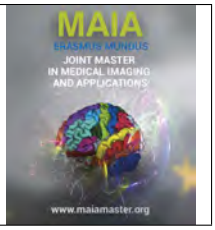
Sanabria, S.J., Goksel, O., 2016. Hand-held sound-speed imaging based on ultrasound reflector delineation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 568–576.

Sanabria, S.J., Goksel, O., Martini, K., Forte, S., Frauenfelder, T., Kubik-Huch, R.A., Rominger, M.B., 2018. Breast-density assessment with hand-held ultrasound: A novel biomarker to assess breast cancer risk and to tailor screening? European radiology , 1–11.

Siu, A.L., 2016. Screening for breast cancer: Us preventive services task force recommendation statement. Annals of internal medicine 164, 279–296.

Medical Imaging and Applications

Master Thesis, June 2018



Transfer learning for automatic detection of Alzheimer's Disease

Katherine Sheran, Mariano Cabezas, Xavier Llado, Arnau Oliver

Computer Vision and Robotics Group, University of Girona, Catalonia, Spain

Abstract

The accurate diagnosis of Alzheimer's disease and its early stage, mild cognitive impairment, is essential for timely treatment and possible delay of the disease. In recent years there has been a great interest in using neuroimaging data in combination with machine learning techniques in order to correctly and promptly diagnose Alzheimer. Recent success of deep learning in computer vision has improved such research and its use is likely to grow rapidly. However, common limitations with these implementations are the requirement of large amounts of labeled training images and multiple processing steps for feature extraction. The present work attempts to solve this issue by using transfer learning, where popular architectures, such as VGG and Inception, are initialized with pre-trained weights from large datasets of natural images and the fully-connected layer is re-trained with only a small number of structural MRI brain scans. The proposed approach consists on preprocessing the MRI volumes, followed by a selection of the most informative slices and finally, retraining the final layer of a network in order to perform both binary classification and multi-class classification. Performance was evaluated for classification of Alzheimer's disease versus mild cognitive impairment and normal controls on the Alzheimer's Disease National Initiative (ADNI) and the Open Access Series of Imaging Studies (OASIS) datasets of 3D structural MRI brain scans. The proposed implementation demonstrates that with a smaller training size and fewer processing steps, comparable or even better performance is reached than current state-of-the-art methods, resulting in high and reproducible accuracy rates.

Keywords: Alzheimer's disease, mild cognitive impairment, magnetic resonance imaging, deep learning, transfer learning, convolutional neural network, computer aided diagnosis

1. Introduction

1.1. Alzheimer's Disease

Alzheimer's Disease (AD) is an irreversible neurodegenerative disease that results in a loss of mental function due to the progressive death of brain cells. It is characterized by a decline in memory, thinking, problem solving and ability to formulate and use language. This decline occurs because neurons in the section of the brain involved in cognitive function are deteriorated and no longer function properly. The damage eventually causes dementia and affect parts of the brain that enable a person to carry out basic bodily functions, such as speaking and swallowing, which can ultimately lead to the individual's death. AD is currently the most common neurodegenerative disorder and the most frequent cause of dementia. It is also considered the main cause of death for people over 65 years old (Brookmeyer et al., 2007). According to the World Alzheimer's report of 2017, there was an estimate of 47 million people worldwide diagnosed with Alzheimer's or a related dementia. The incidence of AD is expected to rise as the population of oldest adults increases due to gains in life expectancy. AD is expected to double every 20 years, reaching 76 million in 2030 and 131.5 million in 2050 (Sabbagh and Decourt, 2017). This means that 1.2% of the world population will be affected by AD, with a major prevalence in Europe and North America. The global cost of Alzheimers and dementia is estimated to be $605 billion, which is equivalent to 1% of the entire world's gross domestic product. By 2050, costs associated with dementia could rise up to $1.1 trillion (Asghar et al., 2017). The impact of AD on patients and their families, the health care system, and society, is enormous and growing, which makes finding effective solutions to reduce the physical, emotional and financial

burdens of this disease a worldwide priority.

Unfortunately, up to this day, no cure has been found for Alzheimer's disease. The current goal of treatment is to slow the progression of the disease and manage its symptoms. Although this is very difficult, it is possible to a certain extent if it is diagnosed relatively early on. However, present research shows that most people living with Alzheimer's have not received a formal diagnosis (Jeon et al., 2017). In high income countries, only 20-50% of dementia cases are recognized, whereas in low and middle income countries, less than 10% of cases are diagnosed (Prince et al., 2013). Even in the latest stages of the disease, diagnosis is inaccurate 50% of the time (Boise et al., 2004).

The problematic nature of diagnosing AD resides on the fact that its symptoms (confusion, memory loss, decreased vision and hearing) also manifest themselves in a normal healthy aging process and in other types of dementia. A healthy aging process may involve a decrease in hearing and vision. It is also common to have a slight decline in memory, however, cognitive decline that impacts daily life is not considered a normal part of aging. Dementia is an overall term that describes a group of symptoms associated with a decline in cognitive skills severe enough to reduce a person's ability to perform everyday activities. It can result from various diseases that cause damage to brain cells, like Parkinson's or Huntington's disease. Hence, it is important for physicians to correctly identify the cause of dementia as it can lead to timely treatment and a delay in morbidity (Petersen et al., 2001).

The early diagnosis of AD is primarily associated to the detection of Mild Cognitive Impairment (MCI), an intermediate stage between the expected cognitive decline of normal aging and the more serious decline of dementia. Figure 1 shows the different evolution of normal aging and dementia. MCI symptoms involve problems with memory, language, thinking and judgment that are greater than normal aging related changes. Although the memory complaints and deficits of MCI do not notably affect the patient's daily activities, it has been reported that individuals with MCI have a high risk of progression to Alzheimer's or other forms of dementia (Gauthier et al., 2006).

Current diagnosis of MCI or AD relies mostly on subjective clinical observations and cognitive tests which include patient history, a mini mental state examination (MMSE), as well as physical and neurobiological exams (Sarraf and Tofighi, 2016). Up to this day, there is no single test that can effectively diagnose AD. According to the Alzheimers Association, AD can only be probable during the patient life, whereas a definite diagnosis requires postmortem histopathological confirmation.

Research efforts are focused on discovering an accurate way of detecting the disease. In recent years, a number of researchers have identified disease specific biomarkers of AD. This reliable identification of biomarkers supports a major change in the diagnosis of dementia as it allows physicians to combine clinical observations with in-vivo biological manifestations and integrate the information into the diagnosis process (Dubois et al., 2010). These identifiable imaging biomarkers have been effectively used for the diagnosis or prognosis of AD due to their advantages of visualization and quantitative measurements by neuroimaging.

Among the different neuroimaging modalities, structural magnetic resonance imaging (sMRI) has been recognized as a promising indicator for the early diagnosis of Alzheimer's Disease and its progression (Bron et al., 2015). AD has a certain progressive pattern of brain tissue damage. It shrinks the cerebral cortex of the brain and enlarges the ventricles. Research suggests that the thickness of the cortex in the brain and the size of the ventricles are representative biomarkers for predicting and diagnosing Alzheimer's Disease (Querbes et al., 2009). Figure 2 displays the aforementioned biomarkers on axial sMRI scans of a normal control (NC), a patient with late MCI and a patient with severe AD.

Another quantifiable biomarker provided by an sMRI scan is the degree of neural degeneration and the affected brain zones. About 5% of neurons die each year in someone with Alzheimers, compared to less than 1% in a senior who is aging normally (O'Kelly, 2016). Since brain cells in the damaged regions have degenerated, they display lower intensities on the sMRI scan. This provides additional information for the diagnostic criteria.

Identifying the visual distinctions between sMRI images of AD, MCI and senior patients with normal aging requires extensive knowledge and experience, which must be also combined with additional clinical observations in order to accurately classify the data. However in some cases this distinction is not so easily noticeable. Early signs of AD are difficult to differentiate from MCI; similarly, scans of healthy aged subjects are difficult to distinguish between an early MCI stage. While promising, brain imaging remains an underutilized resource for aiding medical experts in performing early diagnosis due to limitations of the human eye. Therefore, the development of an assertive and automatic tool for classifying scans between healthy, MCI and AD patients is of great interest to clinicians.

Classification between scans can be achieved through automated analysis of sMRI images with machine learning. Recent studies have shown that machine learning algorithms were able to predict AD more accurately than experience clinicians (Klöppel et al., 2008), making it an important field of research for computer aided diagnosis. Attention in medical imaging has thus been shifted into finding biomarkers and applying machine learning techniques to perform automatic early detection of AD. A multitude of machine learning methods have been tried for this task in recent years, including support vector machines, independent component anal-

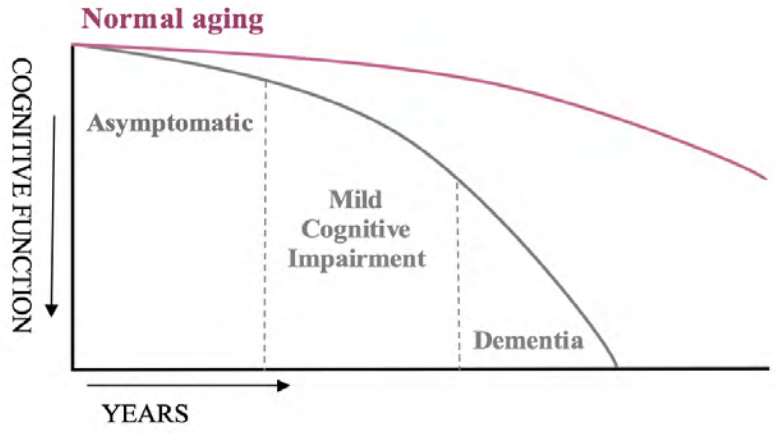


Figure 1: The course of Alzheimer's disease evolves continuously through a stage known as mild cognitive impairment.

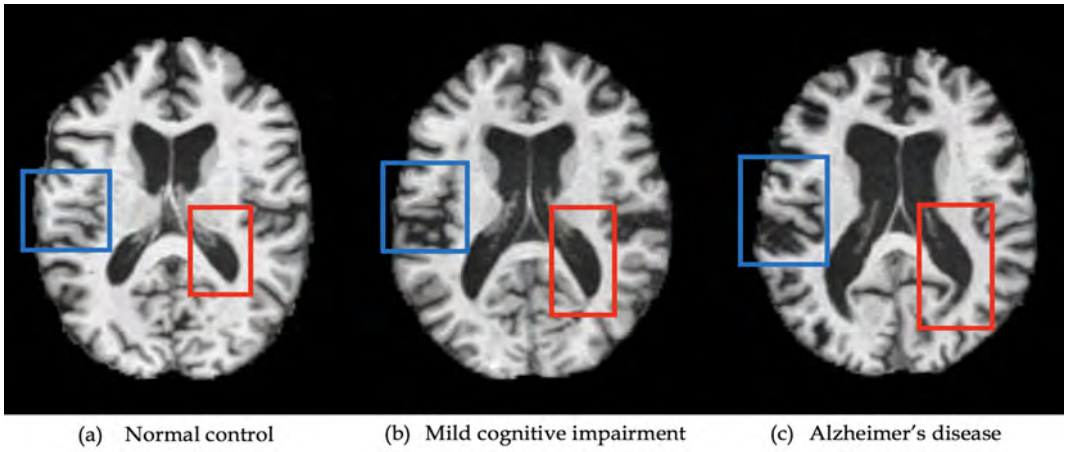


Figure 2: Axial sMRI brain scans of (a) normal control, (b) mild cognitive impairment and (c) Alzheimer's disease subjects from the ADNI dataset. Representative biomarkers for detecting Alzheimer are highlighted: in blue, shrinkage of cerebral cortex; in red, ventricle enlargement.

ysis and penalized regression. While these statistical machine learning methods have proved to be effective in diagnosing AD from neuroimages, recent deep learning methods, such as convolutional neural networks, have outperformed the traditional statistical methods.

1.2. Deep Learning

Deep learning (DL) is a subfield of machine learning, based on learning data representations. Learning can be supervised, in which a label or groundtruth is provided to train the algorithm; or unsupervised, were there is no previous criteria and the computer itself determines the classes of the images. Convolutional neural networks (CNNs or convnets), whose design is roughly similar to the human vision system, are the pillar algorithms of deep learning. They are one of the best models for solving perceptual problems, such as identifying images, classifying and clustering them. CNNs have recently attracted much attention as they have proven to be able to recognize thousands of object categories from natural image databases, as shown on the ImageNet challenge, a benchmark of machine learning performance (Dean et al., 2018). In this annual competition, teams compete to classify millions of images into categories. A milestone year was 2012, when Alex Krizhevsky used the first neural network entry and won the competition by dropping the classification error record from 26% to 15%, an astounding improvement at the time (Zheng et al., 2018). Since then, every winning entry has used a deep learning architecture, with performance now exceeding that of humans (Zaharchuk et al., 2018).

The core of CNNs are convolutional layers which can extract local features (e.g. edges) across an input image through convolution. Each node in a convolutional layer is connected to a small subset of spatially connected neurons. To reduce computational complexity, a max pooling layer follows convolutional layers, which reduces the size of feature maps by selecting the maxi-

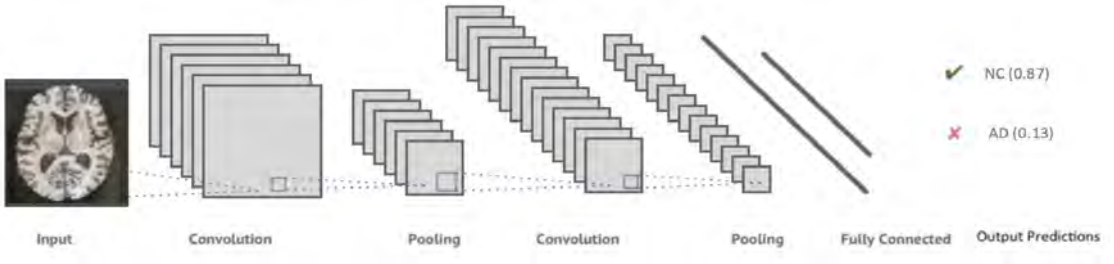


Figure 3: Example of a CNN architecture: an input image, in this case a 2D slice, is processed throughout the network over many layers and the output neurons hold the probabilites for belonging to a class.

mum feature response in a local neighborhood. Pairs of convolutional and pooling layers are followed by a number of fully-connected layers, where a neuron in one layer has connections to all activations in the previous layer. Fully-connected layers help learning non-linear relationship among the local features extracted by convolutional layers. Finally, a soft-max layer follows the fully-connected layers, which normalizes the outputs to desired levels. Figure 3 displays an example of a CNN architecture used for classification. CNNs are trained with the back-propagation algorithm, where in each iteration, weights associated with the neurons in the convolutional layers are updated in a way that minimizes a cost function. When training from scratch, the weights are typically initialized randomly, drawing from a normal distribution.

CNNs have become a popular neural network architecture, with the potential to revolutionize entire industries, including medical imaging analysis. The main power of a CNN lies in its deep architecture which allows the extraction of a set of discriminating features from images. These features help the network identify to which class the image belongs to and assign its corresponding label. However, training a deep CNN from scratch is not a simple task. First, a huge amount of labeled data is required. In the medical domain, this requirement is difficult to meet as obtaining expert annotations is timely, expensive and prone to variability. Diseases might be scarcely represented and data could be protected due to ethical reasons. Second, training a deep CNN requires extensive computational and memory resources. Without the appropriate equipment, the training process becomes extremely time consuming. Third, training requires careful and tedious tuning of many parameters which, done incorrectly can result in over-fitting and convergence issues. Overall, training a network from scratch requires a large amount of labeled training data and a great deal of expertise to ensure proper convergence.

A more attractive alternative to training from scratch is to fine-tune a CNN that has already been trained using a large set of labeled natural images, like the ones provided on the ImageNet database. This pre-trained network would have already learned features that are useful for problems, like image classification. Through transfer learning, these extracted features can be used as a starting point to work in a different and smaller dataset, requiring only to fine tune the final layers. Transfer learning has proven to be robust and faster than training from scratch. It has already been applied in the medical imaging field, with results demonstrating that the use of a pre-trained CNN with adequate fine-tuning could perform as well as a CNN trained from scratch (Tajbakhsh et al., 2016).

In the present master thesis, popular CNN architectures were implemented to solve the AD diagnosis problem by using transfer learning. The first architecture, VGG, is a 16-19 layer network built by Oxford's Visual Geometry Group, hence the name VGG (Simonyan and Zisserman, 2014). The second architecture, Inception V3, is a variant of the GoogLeNet model which was a state of the art image recognition net in 2014 (Szegedy et al., 2017). These architectures are both previous winners of the ImageNet challenge and are open source, meaning that the architectures, as well as the pretrained weights, are readily available online. Although the architectures are trained on a different domain (natural images from the ImageNet database), we demonstrate that with the help of transfer learning, the pre-trained weights of these networks can be adapted to our database of brain sMRI scans.

1.3. Data acquisition

The structural MRI scans used in the present work were obtained from two well-known public dementia datasets. The first one, the Open Access Series of Imaging Studies (OASIS), is a project aimed at making neuroimaging data sets of the brain freely available to the scientific community. OASIS provides crossectional MRI scans of nondemented and demented older adults (Marcus et al., 2007). The second dataset comes from the Alzheimers Disease Neuroimaging Initiative (ADNI) , which unites researchers with study data as they work to define the progression of Alzheimers disease. ADNI researchers collect, validate and utilize data, such as structural MRI images, as predictors for the disease. Study participants include Alzheimers dis-

ease patients, mild cognitive impairment subjects and elderly controls (Mueller et al., 2005).

1.4. Aim and Objectives

The aim of this master thesis is to develop a computer aided diagnosis tool in order to detect Alzheimer's disease and its early stage, mild cognitive impairment, using transfer learning on structural MRI images. In particular, our objectives are:

- Present a state of the art review on the diagnosis of AD based on structural MRI images and machine learning techniques.
- Fine tune pre-trained CNN architectures to perform the difficult diagnostic task of classification between three classes of subjects: NC, MCI and AD patients.
- Identify which of the pre-trained CNN architectures adapts better to our classification problem.
- Implement the developed architectures on different dementia datasets to confirm robustness of the proposed model.
- Compare the proposed deep learning approach with state of the art machine learning methods for AD diagnosis.

1.5. Organization of the document

This document is organized in the following manner. First, a state of the art, background and theory are presented in section 2. Next, in section 3 a description of the materials and methods is provided. In Section 3.2, all the steps of the proposed method are described. In section 4, the results and all relevant findings are presented. The discussion of results and findings is presented in section 5. Finally, section 6 concludes and present future work possibilities.

2. State of the art

2.1. Classical machine learning methods

Several studies have demonstrated that it is possible to automatize the diagnosis of Alzheimers disease using computer aided systems based on machine learning in combination with structural MRI images. Among the many approaches, classical methods, like support vector machines (SVMs), have been extensively used in the area providing good results. An interesting work is the one of Klöppel et al. (2008) who used SVMs with linear kernels for the classification of gray matter signatures and compared their results against the performance achieved by expert radiologists. They demonstrated that their algorithm provided a better diagnosis of Alzheimer than the one of human experts, reaching an accuracy value of 93%, whereas radiologists performance ranged between 80%-90% when discriminating between AD and NC patients.

One recent method using the ADNI dataset is the one of Jha et al. (2017), which uses a dual-tree complex wavelet transform for extracting features from an sMRI. The dimensionality of the feature vector is reduced by using principal component analysis (PCA) and the reduced feature vector is sent to a feed-forward neural network (FNN) in order to distinguish AD and NC subjects, achieving an accuracy of 90.06%.

Another work is the one of Liu et al. (2016), who extracted multi view features using several selected templates from the ADNI dataset. Tissue density maps of each template were used then for clustering subjects within each class in order to extract an encoding feature of each subject. Finally, an ensemble of SVMs were used to perform classification, resulting in accuracy values of 93.83% for AD vs MC and 89.1% for MCI vs NC.

An example of a classical machine learning method using the OASIS dataset is the one of Amulya et al. (2017), who extracted texture features using a Gray-Level Co-occurrence Matrix (GLCM) method and performed classification between AD and NC using SVMs. Obtaining an accuracy value of 75.71%.

2.2. Deep Learning methods

Although classical machine learning methods have proven to be efficient in diagnosing AD, recently deep learning techniques have outperformed these methods by a large margin. Such is the case of Payan and Montana (2015), who designed a predictive algorithm to distinguish AD, MCI and NC subjects by combining a sparse autoencoder with a 3D CNN architecture, obtaining an accuracy value of 95.39% for classifying AD from NC subjects. In this work they also implemented a three way classifier, obtaining an accuracy value of 89.4% for discriminating AD, MCI and NC subjects.

Gupta et al. (2013) employed 2D CNN for slice-wise feature extraction of sMRI scans. To boost the classification performance, the CNN was pretrained using a Sparse Autoencoder (SAE), reaching a final accuracy of 94.7% for detecting AD from NC.

The work of Hosseini-Asl et al. (2016) aimed to extract features related to AD variations of anatomical brain structures, such as ventricles size, cortical thickness and brain volume, using a three dimensional convolutional autoencoder. The autoencoder is pretrained to capture anatomical shape variations in structural brain MRI scans. The encoder is fed into fully connected layers which are then trained for each specific AD classification task. Their experiments on the ADNI dataset have shown better results compared to several conventional classifiers. In addition, they perform three 2-way classifiers, obtaining an overall accuracy of 97.6% for AD vs NC, 95% for AD vs MCI and 90.8% for MCI vs NC classification.

Table 1: State of the art algorithm comparison for automatic classification of sMRI images.

AUTHOR	YEAR	METHOD	DATABASE	NUMBER OF SUBJECTS	CLASSIFICATION ACCURACY (%)			
					AD/NC	AD/MCI	MCI/NC	AD/MCI/NC
Sarraf et al.	2016	LeNet and GoogleNet	ADNI	AD: 211 NC: 91	98.84			
Hossein-Asl et al.	2016	3D-ACNN	ADNI	AD: 70 NC: 70 MCI: 70	97.60	95.00	90.80	
Hon and Khan	2017	Fine-tuning VGG16	OASIS	AD: 100 NC: 100	96.25			
Payan and Montana	2015	Autoencoder and 3D-CNN	ADNI	AD: 755 NC: 755 MCI: 755	95.39	86.80	92.10	89.40
Gupta et al.	2013	2D CNN	ADNI	AD: 200 NC: 232 MCI: 411	94.70	88.10	86.35	85.00
Liu et al.	2016	Multitemplate and SVMs	ADNI	AD: 97 NC: 128 MCI: 234	93.83		89.10	
Islam and Zhang	2017	Deep neural network	OASIS	AD: 100 NC: 135	93.13			
Kloppel et al.	2008	SVMs with linear kernels	Private dataset	AD: 20 NC: 20	93.00			
Jha et al.	2017	Complex wavelet transform	ADNI	AD: 28 NC: 98	90.06			
Amulya et al.	2015	GLCM and SVMs	OASIS	AD: 100 NC: 135	75.71			

The best accuracy to date was obtained in the work of Sarraf et al. (2016), in which popular CNN architectures, such as LeNet and Inception model from Google were used for classifying AD from NC subjects, reaching accuracy values of 98.84% when using sMRI scans from the ADNI database.

The above mentioned studies were all developed using sMRI brain scans from the ADNI dataset. However, some other works have evaluated their performance using the OASIS dataset. This is the case of Islam and Zhang (2017) who developed a deep neural network inspired on the Inception V4 model for AD detection. This deep learning method obtained a final accuracy of 93.12% which is far greater than using conventional machine learning methods. More recently, Hon and Khan (2017), applied transfer learning on pre-trained architectures, such as VGG-16 and Inception V4, achieving an accuracy of 96.25%

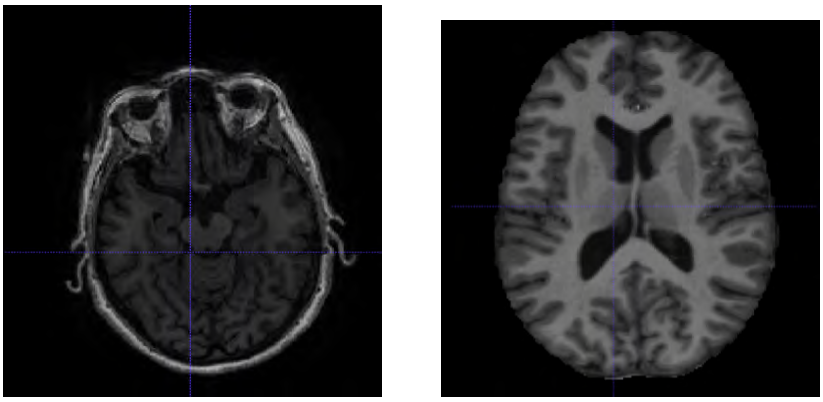
Table 1 contains a summary of the state of the art, describing the implemented methods, the number of subjects used in each study and their reported performance. Although a direct comparison of these studies is difficult, as each study uses different datasets and processing protocols, the table provides a general overview of typical accuracy measures achieved in the classification of sMRI images. In the mentioned studies, a problem of sMRI classification is usually tackled with complex multistage pipelines for feature extraction. Moreover, not all the methods are capable of performing three way classification. In contrast, the present master thesis proposes to develop a deep learning based algorithm that has the potential to simplify the classification pipeline by using transfer learning and significant 2D slices instead of whole MRI volumes. It also tackles the classification problem of classifying subjects between NC, MCI and AD by performing 2-way and 3-way classification.

3. Material and methods

3.1. Materials

3.1.1. ADNI dataset

In this study, the efficiency of the proposed method is evaluated on the Alzheimers Disease Neuroimaging Initiative (ADNI) database available at (http://adni.loni.usc.edu/). There are different data collections of sMRI images: ADNI-1, ADNI-GO, ADNI-2 and ADNI-3. For this work, only raw T1-weighted MRI scans from ADNI-2 were considered. The reason being that ADNI-2 already includes patients from ADNI-1 and ADNI-GO. Moreover, the majority of publications



(a) ADNI-2 raw sMRI (b) OASIS-1 preprocessed sMRI

Figure 4: Sample sMRI volumes from ADNI and OASIS datasets.

Table 2: Statistical data of the ADNI participants.

ADNI SUBJECTS	AD	MCI	NC
No. of patients	100	100	100
Age(years)	75.3±7.4	74.7±7.4	75.8±5.1
Gender(M/F)	67/33	56/44	48/52

Table 3: Statistical data of the OASIS participants.

OASIS SUBJECTS	AD	NC
No. of patients	100	100
Age (years)	75.91±8.98	77.75±6.99
Gender (M/F)	28/72	29/71

report their results using this dataset. Therefore, in order to compare our results with the state of the art, the ADNI-2 data collection was selected. An example of an ADNI volume is shown in Figure 4.a. Since there are patients that have multiple images taken during a period of time, only the baseline images from each subject were selected. Table 2 describes the demographics of the patients in our collection, which is broken into three groups: AD, MCI and NC.

For a balanced dataset, we sampled 100 scans from each group for a total of 300 scans. All the images come in NIfTI-1 format (extension .nii) and include corrections for gradient warping and image inhomogeneity. Further preprocessing was performed (skull stripping, registration and normalization) and the most informative slices from the volume were selected and saved in JPG format for fine tuning the CNN.

3.1.2. OASIS subjects

A second dataset was employed in order to be able to confirm the robustness of the proposed implementation. This dataset consisted on sMRI volumes from the Open Access Series of Imaging Studies (OASIS), accessible at (http://www.oasisbrains.org). OASIS provides two types of data: cross-sectional and longitudinal. Since the aim of this work is to differentiate between AD and NC patients through the brain images, we used the cross-sectional data from OASIS-1. An example of an OASIS volume is shown in Figure 4.b. The dataset consists of 416 subjects whose ages are between 18 and 96. For the experiments, 200 subjects were randomly picked, 100 from the AD group and 100 from the NC group. The demographics of the selected patients are displayed in Table 3. All images come in 16-bit Analyze 7.5 format. No further preprocessing was required as images were already, intensity inhomogeneity corrected, skull stripped, registered and normalized. Only the most informative slices were selected and saved in JPG format for fine tuning.

3.2. Methods

The proposed method consists of three important stages: (1) preprocessing of ADNI-2 volumes, in which skull stripping and affine registration is performed, (2) slice selection, in which central and highest entropy slices are extracted from the volume and (3) classification using transfer learning. As the volumes provided by OASIS-1 were already skull-stripped and registered, they do not require any further preprocessing. Only the slice selection and classification steps apply to this dataset. Figure 5 displays a block diagram of the proposed implementation.

3.2.1. Preprocessing

1. **Skull stripping:** The first preprocessing step applied to the raw sMRI volumes for the ADNI-2 dataset was to remove all non-brain tissue from the images. The Robust Brain Extraction (ROBEX) is a solution to solve the problem of brain extraction from MRI under almost any condition with the benefit of no parameter settings. In the work of

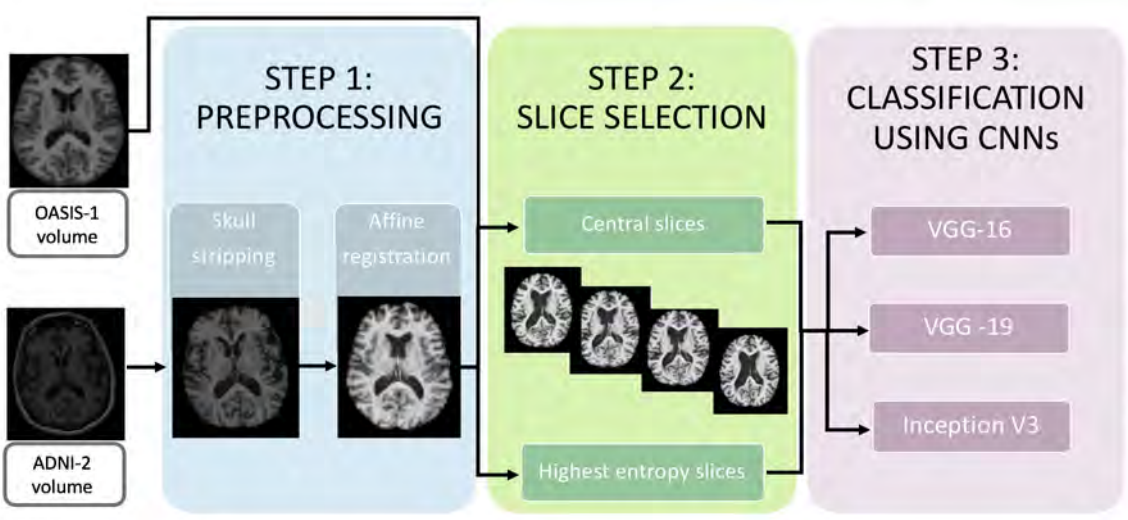


Figure 5: Block diagram of the proposed implementation.

Iglesias et al. (2011), it was demonstrated that ROBEX performed better than six other popular algorithms (FSL-BET, BSE, FreeSurfer, AFNI, BridgeBurner and GCUT). Due to its ease of use, lack of parameters and its good performance, ROBEX was used in this first step.

2. **Image registration:** In the second preprocessing step, FMRIB's Linear Image Registration Tool (FLIRT) was used to perform linear registration applying an affine transformation. In this case, each brain was normalized in shape and pose to the MNI152 standard space. The alignment to the standard space was completed using the affine transformation process provided by FLIRT, which has 12 parameters or degrees of freedom. This process consists of aligning the subject image to the standard space applying 12 parameters, 3 on each sub-group: translation, shear, rotation and scale (Jenkinson and Smith, 2001). The affine transformation was employed to keep the shape and proportions of each brain. This linear registration process helps to compare different substructures independently of the actual size of each brain. FLIRT has several parameters that need to be tuned to perform the registration.

   The parameters used on FLIRT were the following:

   - **bins:** 256 (enabled 256 bins in the intensity histogram).
   - **cost:** corratio (cost function correlation ratio).
   - **dof:** 12 (3 for translation, 3 for shear, 3 for rotation and 3 for scale).
   - **interp:** trilinear (method to estimate the central point of each voxel.
   - **ref:** <path> (MNI152 template filename).
   - **in:** <path> (input image filename).
   - **out:** <path> (registered image filename).

3.2.2. Slice selection

The second major step on this work was to select the most informative slices from the volume. A 3D sMRI volume is composed by several slices that may not contain useful information for detecting AD and might mislead the network during training. This is the case of the first and last slices of the volume, which consist mostly of background. In order to provide the best possible data for training, slices that contain the most relevant information of the brain tissue need to be selected. In the present work, this is done by means of two methods: (1) selecting central slices and (2) selecting the slices with the highest entropy. The entropy provides a measure of variation in an image. Hence, the images with the highest entropy values can be considered as the most informative slices of the volume. The reason behind picking the highest slice from all slices is that it retains more relevant information about the brain tissues as compared to earlier slices and later slices in the volume. The possible direction of the slices are sagital, coronal, and axial. In this research, the axial view was chosen as it is the view that reports the highest accuracy values for AD detection (Glozman and Liba, 2016).

1. **Central slices:** 31 axial slices were selected and extracted from the center of the volume. The slices were saved in JPG format and, in order to be compatible with the pre-trained models of VGG and Inception, the images were resized to be 150x150 for VGG, and 299x299 for Inception. The same process is repeated and applied to all the subjects from both datasets, OASIS and ADNI, resulting in

a total of 6200 images for OASIS and 9300 images from ADNI,

2. **Highest entropy slices:** The entropy values of all slices in the sMRI volume were computed. The 31 axial slices with the highest entropy values were extracted from the volume. The selected slices were saved in JPG format and resized to be 150x150 for VGG, and 299x299 for Inception. The same process is repeated and applied to all the subjects from both datasets, OASIS and ADNI, resulting in a total of 6200 images for OASIS and 9300 images from ADNI.

Table 4, shows a summary of the number of volumes, slices and images used from each database. Experiments were performed with both central slices and the highest entropy slices in order to compare the performance and determine which of the two methods achieve higher accuracy values.

3.2.3. Network architecture and training

The data was randomly divided into 80% training and 20% validation sets on the subject level, rather than on the slice level, to avoid possible bias of having data of the same subject at both the training and testing sets. More details are provided in table 5.

To increase the amount of data available to the network, mirrored images flipped horizontally were added to the training set. Other methods of data augmentation, like scaling, rotating or random cropping were avoided in order to preserve the diagnostic value of the images.

Due to the limited amount of training data, pre-trained networks were used instead of training from scratch. All layers were initialized with ImageNet weights and biases. Only the last fully-connected (FC) layers of the networks were replaced in order to work with our three classes (AD, MCI, NC), instead of the original 1000 ImageNet classes. The previous layers remained frozen because the existing weights are valuable at finding and summarizing features that are useful for image classification problems. Earlier layers in the model contain general features, such as edge or shape detectors, which are ubiquitous for many image classification tasks, whereas the last layers contain higher level features (Weng et al., 2017). This not only provides a robust set of pre-trained weights to work with, but it also allows the use of different architectures to test the power of transfer learning. The following CNN architectures, loaded with pre-trained ImageNet weights, were used in the present work:

1. **VGG-16:** This network is characterized by its simplicity, using only 33 convolutional layers stacked on top of each other in increasing depth. The "16" stands for the number of weight layers in the network. Reducing volume size is handled by max pooling. Two fully-connected layers, each with 4,096 nodes are then followed by a softmax classifier (Simonyan and Zisserman, 2014). For our transfer-learning implementation, the architecture was frozen up until the last convolutional part and the original fully connected layers were replaced with three new ones: (1) a dense layer with a ReLu activation function (2) a dropout of 0.03 and (3) a softmax classifier. Figure 6 displays the frozen and the new FC layers implemented on VGG-16. Training was done with a batch size of 40, a learning rate of 0.001 and using an RMSprop optimization model. An early stopping strategy was adopted to monitor the validation accuracy with a patience set to 5 epochs.

2. **VGG-19:** Made by 19 layers using small convolution filters of size 3x3. As in the previous case, a classifier model consisting of three fully connected layers was built: (1) a dense layer with a ReLu activation function (2) a dropout layer of 0.03 and (3) a softmax classifier. The layers were frozen up to the last convolutional block and the three new built FC layers were added on top of the frozen architecture. The resulting architecture is shown in Figure 7. Training was done with a batch size of 40, a learning rate of 0.001 and using an RMSprop optimization model. Training was stopped when the accuracy did not improve after 5 epochs.

3. **Inception V3:** A variant of deep learning architecture built by Google, made of 22 layers with a 4 parallel pathway of 1x1, 3x3 and 5x5 convolutions. The architecture allows the model to recover both local features via smaller convolutions and high abstracted features via large ones (Szegedy et al., 2016). For our transfer-learning implementation, the first layers remained unchanged, whereas the last FC layers were replaced with: (1) flatten layer, (2) dense layer with ReLu activation, (3) dropout of 0.2, (4) dense layer with ReLu activation and (5) a softmax classifier. Figure 8 displays the final configuration of the model. Training was done with a batch size of 8 and an Adam optimizer with a learning rate of 0.01. Training was stopped when the accuracy did not improve after 10 epochs.

The input of these CNNs models are the sMRI slices, and the output is the probability distribution over the three classes. The slice is labeled according to the largest probability class. The predicted label is then assigned to the patient by majority voting of all slices.

3.3. Evaluation metrics

The performance of disease classification is evaluated by computing the accuracy and the confusion matrix on the obtained predictions from the CNNs.

Table 4: Number of patients, slices and images from OASIS and ADNI datasets used for training and validation.

Class	OASIS DATASET			ADNI DATASET		
	No.patients	No.slices	No.images	No.patients	No.slices	No.images
AD	100	31	3100	100	31	3100
MCI	-	-	-	100	31	3100
NC	100	31	3100	100	31	3100
TOTAL	200	31	6200	300	31	9300

Table 5: Statistics on training and validation data.

Class	OASIS DATASET		ADNI DATASET	
	Training	Validation	Training	Validation
AD	2480	620	2480	620
MCI	-	-	2480	620
NC	2480	620	2480	620
Total no. slices	4960	1240	7440	1860

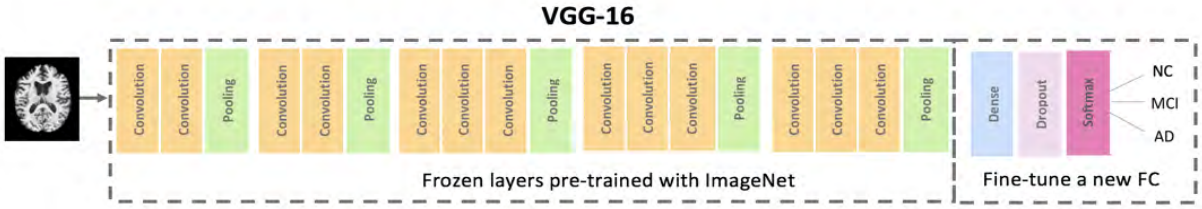


Figure 6: VGG-16 transfer learning layout.

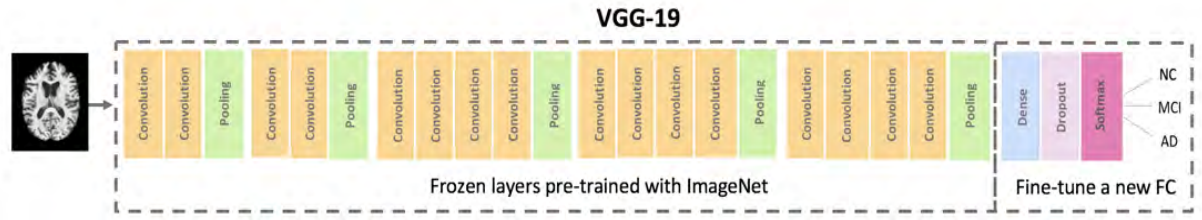


Figure 7: VGG-19 transfer learning layout.

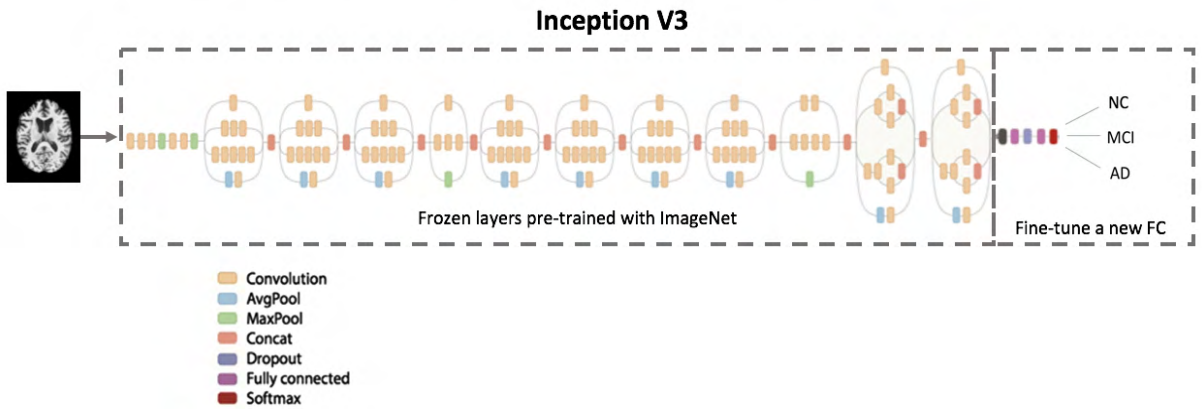


Figure 8: Inception V3 transfer learning layout.

3.3.1. Confusion matrix

The performance is calculated on the essence of the overall confusion matrix, which holds the correct and incorrect classification results.

3.3.2. Accuracy

The accuracy is one common empirical measure to access effectiveness of a classifier. It is calculated as the sum of correct classifications, divided by the total number of classifications, as shown in Equation 1:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

3.4. Software environment

The proposed deep learning implementation was developed on Python using Keras with Tensorflow backend. Keras Applications contained the CNN architectures used in this work, which are VGG-16, VGG-19 and Inception V3. ImageNet weights were downloaded automatically when instantiating a model. The implementation was developed on Ubuntu 14.04 running an Intel® Core ™ i7, with a processing speed of 3.60GHz and 31GB of RAM. Training was performed with a TITAN X (Pascal) GPU.

4. Results

In this section, the experimental results are presented for both OASIS and ADNI datasets. For the OASIS images, only 2-way classification was performed (AD vs NC) as the database does not provide the MCI class. Whereas for the ADNI dataset, classification results are presented using three binary classifiers (AD vs NC, AD vs MCI and NC vs MCI), and one 3-way classifier (AD vs NC vs MCI). The experiments were mainly divided in two sections: (1) using central slices and (2) using highest entropy slices. A 5-fold cross-validation strategy was used to evaluate the performance of the method. The average accuracy of all 5-folds is computed and reported for each of the implemented architectures. For a better visualization of the multi-class problem, a confusion matrix and examples of correctly and incorrectly classified slices are displayed. Finally, a comparison between the proposed method and the state of the art is performed.

4.1. OASIS dataset

4.1.1. 2-way classification

The first set of experiments were performed with the OASIS dataset. As the data was already preprocessed, it was faster to implement into our transfer-learning layout and to fine-tune the parameters for further experiments with the ADNI dataset. Table 6 shows the individual accuracy of VGG-16, VGG-19 and Inception V3 architectures for a 2-way classification between AD and NC on the OASIS dataset. The highest accuracy value was 99.84% obtained from fine-tuning VGG-16 on the slices with highest entropy. Table 7 shows the comparison of the proposed implementation against the state of the art techniques in terms of accuracy and training size on the OASIS dataset. The comparison demonstrates that the proposed method surpasses all of the previous works in terms of accuracy.

4.1.2. ADNI dataset

4.2. 3-way classification

After obtaining good performance with the OASIS dataset, we decided to begin our tests with the preprocessed ADNI slices. The first experiment performed with ADNI was to implement the 3-way classification between AD, NC and MCI. Table 8 shows the individual performance of VGG-16, VGG-19 and Inception V3 architecture for a 3-way classification between AD, MCI and NC on the ADNI-2 dataset. The highest accuracy value was 66 .23% obtained from fine-tuning Inception-V3 on the highest entropy slices. For a better understanding of the multi-class problem, the confusion matrix from Inception V3 is displayed in Figure 9. In addition, examples of correctly and incorrectly classified slices from Inception V3 are shown in Figure 10. As the performance was not as good as the one obtained with OASIS, we decided to simplify the 3-way classification task into three binary and simpler classifications in order to achieve better results.

4.3. 2-way classification

Table 9 reports the individual performance of VGG-16, VGG-19 and Inception V3 models respectively, for a three binary classification: AD vs NC, AD vs MCI and NC vs MCI. For discriminating AD vs NC, the best accuracy was 98.52%, obtained using VGG-16 on the highest entropy slices. The highest accuracy obtained for classifying AD vs MCI subjects was 75.44% obtained from fine-tuning Inception V3 on the highest entropy slices. Finally, for classifying NC vs MCI subjects, the best accuracy was 82.27% once again obtained from using Inception V3 and the highest entropy slices. In order to know if the obtained accuracy values were above average, we compare all the classification tasks with the state of the art. Table 10 shows the comparison between the proposed implementation and the state of the art methods working on the ADNI dataset. The comparison demonstrates that our implementation outperforms the majority of the techniques when classifying between AD and NC. Reaching comparable results as the ones obtained from Sarraf and Tofighi (2016) which currently hold the best performance. However, when introducing the MCI class, our performance drops drastically. The obtained results were only able to surpass the implementation proposed by Liu et al. (2016).

Table 6: Tested models and corresponding average accuracy(%) from 5-fold cross validation on the OASIS dataset for a binary classification between AD and NC subjects.

Slice	Class	OASIS accuracy (%)		
		VGG-16	VGG-19	InceptionV3
Central	ADvsNC	98.68	96.51	93.66
Entropy	ADvsNC	**99.84**	97.62	95.43

Table 7: Comparison with the state of the art in terms of accuracy and training size on the OASIS dataset.

AUTHOR	YEAR	METHOD	NUMBER OF SUBJECTS	ACCURACY (%) AD/NC
Hon and Khan	2017	Fine-tuning Inception V4	AD: 100 NC: 100	96.25
Islam and Zhang	2017	Deep neural network	AD: 100 NC: 135	93.13
Amulya et al.	2015	GLCM and SVMs	AD: 100 NC: 135	75.71
Proposed method	2018	Fine-tuning VGG-16	AD: 100 NC: 100	99.84

Table 8: Tested models and their corresponding average accuracy (%) from 5-fold cross validation on the ADNI dataset for a 3-way classification between AD, MCI and NC.

Slice	Class	ADNI accuracy (%)		
		VGG-16	VGG-19	Inception V3
Central	AD vs MCI vs NC	50.32	61.16	63.67
Entropy	AD vs MCI vs NC	51.88	62.23	**66.23**

Table 9: Average accuracy (%) achieved through across 5 folds on the ADNI dataset for three binary classifications.

Slice	Class	ADNI accuracy (%)		
		VGG-16	VGG-19	Inception V3
Central	AD vs NC	98.24	98.04	91.21
	AD vs MCI	74.36	55.79	65.16
	NC vs MCI	77.87	69.54	64.93
Entropy	AD vs NC	**98.52**	96.11	91.56
	AD vs MCI	65.74	68.77	**82.27**
	NC vs MCI	78.57	70.14	**75.44**

5. Discussion

Transfer learning was performed on three different architectures: VGG-16, VGG-19 and Inception V3 on two different datasets, OASIS and ADNI, with two different sets of inputs, central slices and highest entropy slices. The reported results demonstrate that fine-tuning a network with a small dataset can be used instead of training a network from scratch.

When comparing the performance of each network, we can identify that the VGG-16 model resulted in a higher level of accuracy than VGG-19 and Inception V3 when classifying between AD and NC subjects, with the highest overall accuracy rate of 99.84% for OASIS and 98.52% for ADNI. However, when performing a 3-way classification between AD, NC and MCI, the architecture that provided a better performance was Inception V3, followed by VGG-19 and lastly VGG-16, with accuracy values of 66.23%, 62.23% and 51.88% respec-

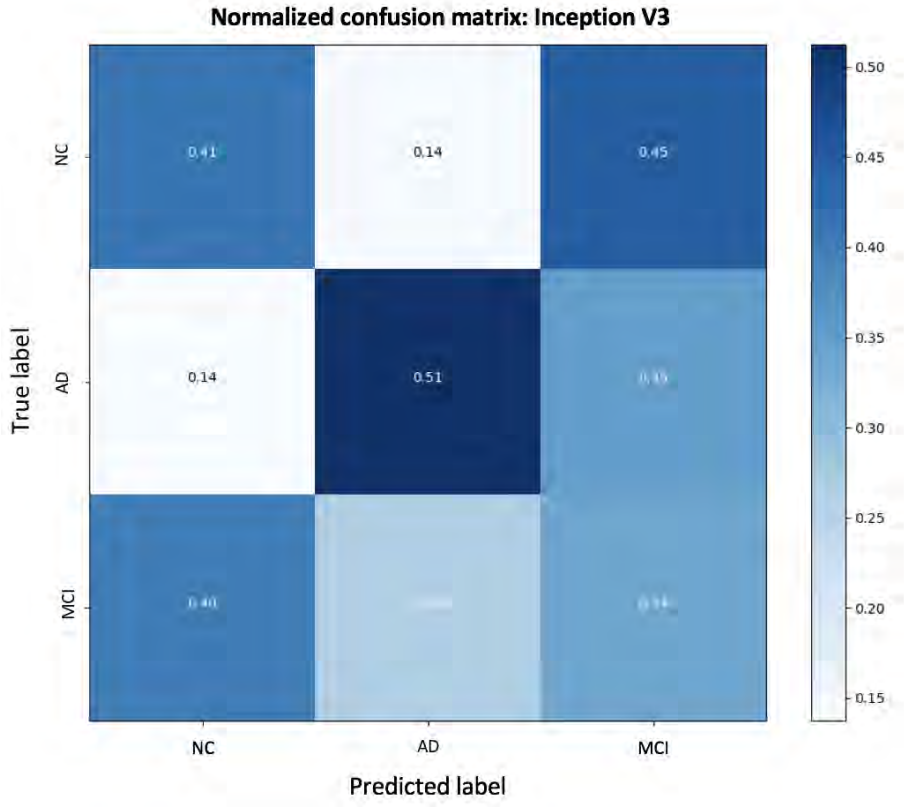
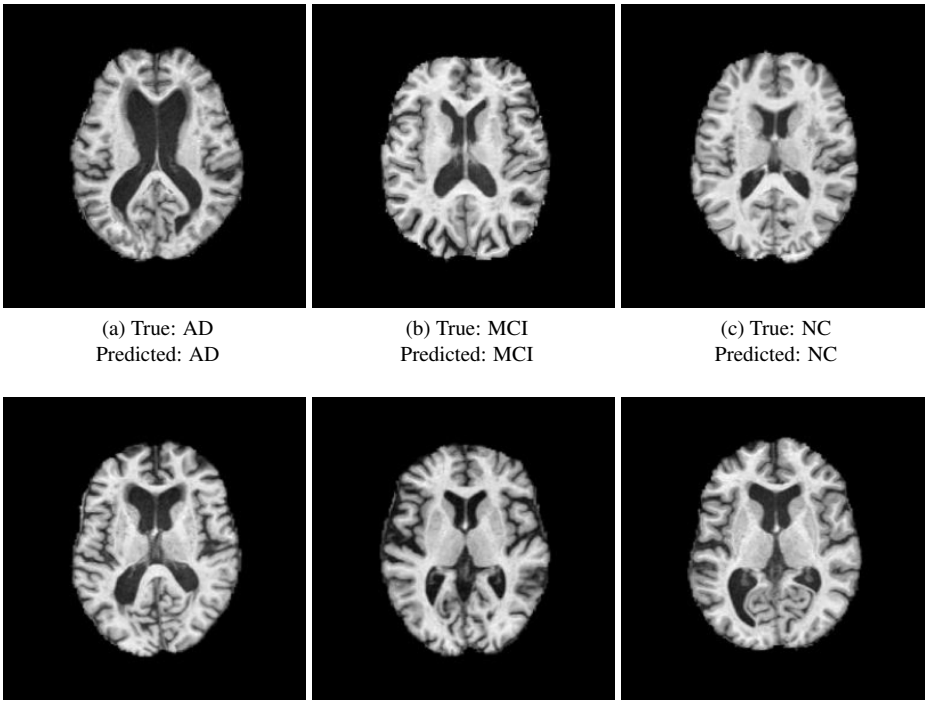


Figure 9: Normalized confusion from Inception V3 for mulit-class classification of ADNI dataset.



(a) True: AD
Predicted: AD

(b) True: MCI
Predicted: MCI

(c) True: NC
Predicted: NC

(d) True: AD
Predicted: MCI

(e) True: MCI
Predicted: NC

(f) True: NC
Predicted: MCI

Figure 10: Example of correct (top) and incorrect (bottom) classification of slices by Inception V3.

Table 10: Comparison with the state of the art in terms of accuracy and training size on ADNI dataset.

AUTHOR	YEAR	METHOD	NUMBER OF SUBJECTS	CLASSIFICATION ACCURACY (%)			
				AD/NC	AD/MCI	MCI/NC	AD/MCI/NC
Sarraf et al.	2016	LeNet and GoogleNet	AD: 211 NC: 91	98.84			
Hossein-Asl et al.	2016	3D-ACNN	AD: 70 NC: 70 MCI: 70	97.60	95.00	90.80	
Payan and Montana	2015	Autoencoder and 3D-CNN	AD: 755 NC: 755 MCI: 755	95.39	86.80	92.10	89.40
Gupta et al.	2013	2D-CNN	AD: 200 NC: 232 MCI: 411	94.70	88.10	86.35	85.00
Liu et al.	2016	Multitemplate and SVMs	AD: 97 NC: 128 MCI: 234		70.10	77.40	
Kloppel et al.	2008	SVMs with linear kernel	AD: 20 NC: 20	93.00			
Jha et al.	2017	Complex wavelet transform	AD: 28 NC: 98	90.06			
Proposed method	2018	Fine-tuning VGG-16	AD: 100 NC: 100 MCI: 100	98.52	65.74	78.57	51.88
Proposed method	2018	Fine-tuning Inception V3	AD: 100 NC: 100 MCI: 100	91.56	75.44	82.27	66.23

tively. In a similar manner, when performing a binary classification between AD vs MCI, Inception V3 outperformed the rest of the networks, with an accuracy of 82.27%. The same happened when classifying NC vs MCI subjects, with Inception V3 giving the highest accuracy of 75.44%.

The different performance in the architectures might be due to the complexity of the problem and the architecture of the models. When performing a simpler task, which is classifying AD against NC, VGG-16 and VGG-19 perform well. However, when introducing a third class, the MCI class, which visually is very similar to the AD and NC subjects, the problem increases greatly in complexity. As mentioned previously, Inception V3 uses parallel pathways with different sizes of convolutions. The architecture allows the model to recover local features by using smaller convolutions and high abstract features by using larger ones. This ability to extract different types of features may be the reason why it outperforms VGG-16 when classifying MCI subjects.

The confusion matrix displayed in Figure 9 provides us a better understanding of the multi-class problem. In this image we can observe clearly how the network confuses the MCI class with AD and NC subjects. The probability of MCI of correctly being classified as MCI is only of 34%. Whereas, in the majority of the cases MCI is actually being classified as NC with a probability of 40%. A similar thing occurs when classifying NC subjects. The majority of NC cases are actually being classified as MCI patients, with a probability of 45% and a probability of being correctly classified as NC of only 41%. On the other hand, the AD class is the only that is being correctly identified by the network, with a probability of 51% of being correctly classified, and a probability of 0.35% of being considered as an MCI case.

The reason behind this misclassification can be comprehended when looking at Figure 10, in which examples of correct and incorrect classifications are displayed. The difference in ventricle size is usually bigger between AD subjects and the rest of the classes. However, MCI and NC subjects have similar features that

might confuse the network. One way to solve this problem might be to provide additional information to the network, like clinical data or other type of neuroimaging data, like fMRI or PET. In this way the network could have more elements to perform a more accurate diagnosis.

Another way to solve the difficult multi-class classification problem is to reduce its complexity, just like we did in the present work by performing three binary classifications. As shown in Table 9, we obtained a better performance when classifying AD vs MCI, with an overall accuracy of 82.27% when performing on Inception V3. Whereas for the classification of NC vs MCI, the highest value was 75.44%, once again on Inception V3. These results, although not perfect, increased considerably regarding the previous experiment. With a better fine tuning and with additional data, accuracy values are expected to increase.

Regarding the experiments with central and highest entropy slices, all of the reported results demonstrated that the highest entropy slices obtained a higher performance. This means that slices containing key biomarkers for AD detection are not necessarily always positioned in the center of the brain volume. Hence, the highest entropy slice selection is a good approach for selecting the most informative images for performing AD classification.

Reported results demonstrate that OASIS dataset gives a slightest better performance than ADNI, probably because the images from OASIS were already preprocessed. This means that a more detailed preprocessing on the ADNI images could improve the accuracy values. Another reason of decreased performance is the fact that the ADNI dataset comes from different institutes and from different MRI scanners, unlike OASIS in which data is more uniform. This difference in data acquisition could pose an explanation behind the reduced accuracy in ADNI.

Comparable or even better performance than most state of the art methods was achieved when classifying AD patients from NC. As shown in Table 10. The results obtained from fine-tuning VGG-16 managed to surpass all of the methods, except the one of Sarraf and Tofighi (2016), who obtained an accuracy of 98.84% which is slightly better than our accuracy of 98.52%. When performing a 3-way classification we could not surpass the works of Payan and Montana (2015) and Gupta et al. (2013) who obtained accuracy values of 89.40% and 85.00% respectively, whereas our accuracy was only of 66.23% when performing fine-tuning with Inception V3. When performing three binary classifications, the accuracy values improved but results comparable to those of the state of the art were not achievable. We were only able to surpass the accuracy obtained by Liu et al. (2014) which was 70.10% for ADvsMCI and 77.40% for MCIvsNC compared to our 75.44% for ADvsMCI and 82.27% for MCIvsNC.

6. Conclusions

Deep learning methods have become prevalent in computer vision applications. In medical imaging, due to a much smaller amount of labeled data, these techniques face many challenges. An alternative to training a CNN from scratch is to fine tune only the final layers of a network that has already been trained. In this paper, a transfer learning based method is proposed in order to identify AD, MCI and NC patients from structural MRI images. Popular architectures like VGG-16, VGG-19 and Inception V3 were tested, with VGG-16 providing the highest accuracy values when performing classification between AD and NC subjects, whereas Inception V3 provided a better performance when introducing the MCI class into the classification problem.

In this work two datasets, OASIS and ADNI, were used for the first time to compare the performance of a deep learning approach for Alzheimer's disease classification. The obtained results demonstrate the robustness of transfer learning, as high accuracy values were obtained for both datasets when distinguishing AD from NC subjects. The performed experiments demonstrate that higher accuracies were obtained for the OASIS dataset.

A novelty of this approach was the use of an entropy based technique in order to select the most informative slices instead of the whole volume for fine-tuning a CNN. The pre-trained CNN architectures were tested on images from ADNI and OASIS datasets, where slices were extracted from sMRI scans and then used to fine-tune the models.

In this work a state of the art review on the diagnosis of AD based on structural MRI images and machine learning techniques was presented. As reported on this work, the proposed approach provides performance comparable to state of the art despite having fewer processing steps and a smaller training set.

Reported results suggest that with the available data, the network can successfully learn to classify two classes AD vs NC. However, when faced with a three-way classification task, the proposed approach does not achieve good accuracy. The reason for this is not only the limited amount of data but also the ambiguity of it, as the MCI images are very similar to the AD and NC classes.

Future improvements need to be done regarding the three way classification problem. Possible solutions could be to increase the number of data and integrate clinical information about the patient into the network. Another possible approach is to implement modality fusion between PET, fMRI and sMRI images in order to provide more information to the network and increase its diagnosing accuracy.

7. Acknowledgments

This thesis marks the conclusion of my master in the Erasmus Mundus Joint Master Degree in Medical Imaging and Applications (MAIA). It has been two very interesting years, and I have learned much during this challenging, but joyful time. I would like to thank the European Commission for their financial support during this two years. Thank you for giving me the opportunity to be here and to continue my studies.

My deepest appreciation goes to my supervisors, Dr. Xavier Llado and Dr. Arnau Oliver, for helping me do my best and offering their constant guidance and expertise in the development of this work.

Special thanks goes to Dr. Mariano Cabezas for his patience, support, and expert suggestions throughout the development of my work. I would also like to thank the rest of my colleagues in the Computer Vision and Robotics group at the University of Girona, I am really grateful for their assistance, helpful comments and encouragement.

Finally, I would like to thank the collaborators in ADNI and OASIS for their great efforts and willingness to share their data, without which the realization of this work would have not been possible.

References

Amulya, E., Varma, S., Paul, V., 2017. Classification of brain mr images using texture feature extraction. International Journal of Computer Science and Engineering 5, 1722–1729.

Asghar, I., Cang, S., Yu, H., 2017. Assistive technology for people with dementia: an overview and bibliometric study. Health Information & Libraries Journal .

Boise, L., Neal, M.B., Kaye, J., 2004. Dementia assessment in primary care: results from a study in three managed care systems. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences 59, M621–M626.

Bron, E.E., Smits, M., Van Der Flier, W.M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J.M., Steketee, R.M., Orellana, C.M., Meijboom, R., et al., 2015. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural mri: the caddementia challenge. NeuroImage 111, 562–579.

Brookmeyer, R., Johnson, E., Ziegler-Graham, K., Arrighi, H.M., 2007. Forecasting the global burden of alzheimers disease. Alzheimer's & dementia: the journal of the Alzheimer's Association 3, 186–191.

Dean, J., Patterson, D., Young, C., 2018. A new golden age in computer architecture: Empowering the machine-learning revolution. IEEE Micro 38, 21–29.

Dubois, B., Feldman, H.H., Jacova, C., Cummings, J.L., DeKosky, S.T., Barberger-Gateau, P., Delacourte, A., Frisoni, G., Fox, N.C., Galasko, D., et al., 2010. Revising the definition of alzheimer's disease: a new lexicon. The Lancet Neurology 9, 1118–1127.

Gauthier, S., Reisberg, B., Zaudig, M., Petersen, R.C., Ritchie, K., Broich, K., Belleville, S., Brodaty, H., Bennett, D., Chertkow, H., et al., 2006. Mild cognitive impairment. The Lancet 367, 1262–1270.

Glozman, T., Liba, O., 2016. Hidden cues: Deep learning for alzheimers disease classification cs331b project final report.

Gupta, A., Ayhan, M., Maida, A., 2013. Natural image bases to represent neuroimaging data, in: International Conference on Machine Learning, pp. 987–994.

Hon, M., Khan, N., 2017. Towards alzheimer's disease classification through transfer learning. arXiv preprint arXiv:1711.11117 .

Hosseini-Asl, E., Gimel'farb, G., El-Baz, A., 2016. Alzheimer's disease diagnostics by a deeply supervised adaptable 3d convolutional network. arXiv preprint arXiv:1607.00556 .

Iglesias, J.E., Liu, C.Y., Thompson, P.M., Tu, Z., 2011. Robust brain extraction across datasets and comparison with publicly available methods. IEEE transactions on medical imaging 30, 1617–1634.

Islam, J., Zhang, Y., 2017. An ensemble of deep convolutional neural networks for alzheimer's disease detection and classification. arXiv preprint arXiv:1712.01675 .

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Medical image analysis 5, 143–156.

Jeon, Y.H., McKenzie, H., Krein, L., Flaherty, I., Gillespie, J., 2017. Preferences, choices and decision-making: A qualitative study on the experiences of family carers and people with advanced dementia. Alzheimer's & Dementia: The Journal of the Alzheimer's Association 13, P1248–P1249.

Jha, D., Kim, J.I., Kwon, G.R., 2017. Diagnosis of alzheimers disease using dual-tree complex wavelet transform, pca, and feed-forward neural network. Journal of healthcare engineering 2017.

Klöppel, S., Stonnington, C.M., Barnes, J., Chen, F., Chu, C., Good, C.D., Mader, I., Mitchell, L.A., Patel, A.C., Roberts, C.C., et al., 2008. Accuracy of dementia diagnosisa direct comparison between radiologists and a computerized method. Brain 131, 2969–2974.

Liu, M., Zhang, D., Adeli, E., Shen, D., 2016. Inherent structure-based multiview learning with multitemplate feature representation for alzheimer's disease diagnosis. IEEE Transactions on Biomedical Engineering 63, 1473–1482.

Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R., Feng, D., 2014. Early diagnosis of alzheimer's disease with deep learning, in: Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on, IEEE. pp. 1015–1018.

Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2007. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. Journal of cognitive neuroscience 19, 1498–1507.

Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C.R., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L., 2005. Ways toward an early diagnosis in alzheimers disease: the alzheimers disease neuroimaging initiative (adni). Alzheimer's & dementia: the journal of the Alzheimer's Association 1, 55–66.

O'Kelly, N., 2016. Use of machine learning technology in the diagnosis of Alzheimers disease. Ph.D. thesis. Dublin City University.

Payan, A., Montana, G., 2015. Predicting alzheimer's disease: a neuroimaging study with 3d convolutional neural networks. arXiv preprint arXiv:1502.02506 .

Petersen, R.C., Stevens, J.C., Ganguli, M., Tangalos, E.G., Cummings, J., DeKosky, S., 2001. Practice parameter: Early detection of dementia: Mild cognitive impairment (an evidence-based review) report of the quality standards subcommittee of the american academy of neurology. Neurology 56, 1133–1142.

Prince, M., Bryce, R., Albanese, E., Wimo, A., Ribeiro, W., Ferri, C.P., 2013. The global prevalence of dementia: a systematic review and metaanalysis. Alzheimer's & dementia: the journal of the Alzheimer's Association 9, 63–75.

Querbes, O., Aubry, F., Pariente, J., Lotterie, J.A., Démonet, J.F., Duret, V., Puel, M., Berry, I., Fort, J.C., Celsis, P., et al., 2009. Early diagnosis of alzheimer's disease using cortical thickness: impact of cognitive reserve. Brain 132, 2036–2047.

Sabbagh, M., Decourt, B., 2017. Current and emerging therapeutics in ad. Current Alzheimer Research 14, 354–355.

Sarraf, S., Tofighi, G., 2016. Classification of alzheimer's disease using fmri data and deep learning convolutional neural networks. arXiv preprint arXiv:1603.08631 .

Sarraf, S., Tofighi, G., et al., 2016. Deepad: Alzheimer s disease classification via deep convolutional neural networks using mri and fmri. bioRxiv , 070441.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning., in: AAAI, p. 12.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826.

Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE transactions on medical imaging 35, 1299–1312.

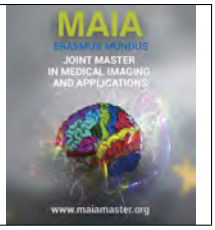
Weng, S., Xu, X., Li, J., Wong, S.T., 2017. Combining deep learning and coherent anti-stokes raman scattering imaging for automated differential diagnosis of lung cancer. Journal of biomedical optics 22, 106017.

Zaharchuk, G., Gong, E., Wintermark, M., Rubin, D., Langlotz, C., 2018. Deep learning in neuroradiology. American Journal of Neuroradiology .

Zheng, Y., Huang, J., Chen, T., Ou, Y., Zhou, W., 2018. Processing global and local features in convolutional neural network (cnn) and primate visual systems, in: Mobile Multimedia/Image Processing, Security, and Applications 2018, International Society for Optics and Photonics. p. 1066809.

Medical Imaging and Applications

Master Thesis, June 2018



Visual Question Answering for Diabetic Retinopathy Screening

Vu Hoang Minh, Raphael Sznitman

University of Bern, ARTORG Center for Biomedical Engineering Research, Murtenstrasse 50, 3008 Bern

Abstract

Recently, researchers from both the computer vision and the natural language processing communities have dramatically shaped queries on a challenging task - Visual Question Answering (VQA). Given an image and a question in natural language, it requires reasoning over visual elements of the image and general knowledge to provide the correct answer.

In this thesis, we propose a fusion scheme with attention network, Weighted Multimodal Low-rank Bilinear Attention Network (WMLB), that outperform the state-of-the-art VQA bilinear models in the most common global VQA dataset. We then transfer our work to the very first VQA methodology in the retinal domain to tackle Diabetic Retinopathy (DR) which is the main source of visual impairment in adults aged 20-74 years. We compare our WMLB with some of the bilinear models providing our state-of-the-art results in both global and retinal domains. We will make our work publicly available.

Keywords: retinal visual question answering, visual question answering, diabetic retinopathy screening

Contents

Figure 1: Examples from balanced VQA v1 dataset. Image was taken from Goyal et al. (2017)

1. Introduction

This thesis aims to investigate the VQA for Diabetic Retinopathy Screening (DRS). The problems are identified in two main areas: VQA and its applications in DRS.

The structure of this section is as follows. Firstly, an overview of the VQA is provided in Section 1.1. The motivation of the project for DRS is then explored in Section 1.2. Next, the objectives of the thesis are described in depth in Section 1.3. Finally, the structure of the document is discussed in Section 1.4.

1.1. Visual Question Answering

The importance of computer science has proliferated in recent decades. This comes from the fact that computer science is a crucial part shaping industry and society with the intent of improving people's quality of life. Machine vision is one field of computer science. It aims at extracting and analyzing useful information from images and videos for industrial and medical applications, for example, medical imaging (Meyer-Baese and Schmid, 2014), signature identification (Sulong et al., 2014), optical character recognition (Trier et al., 1996), handwriting recognition (Xu et al., 1992), object recognition (Lowe, 1999), pattern recognition (Fukunaga, 2013), Simultaneous Localization and Mapping (SLAM) (Engelhard et al., 2011) and the like.

Natural-language processing (NLP) is another field of computer science studying how computers and human languages interact; in particular, how to utilize the power of computers to learn the significant volumes of human language data. As technology is moving very fast and growingly making the communication platforms more approachable, NLP will become an essential technology in bridging the gap between human communication and digital data. The number of NLP applications is steadily increasing like machine translation (Koehn et al., 2007), fighting spam (Heymann et al., 2007), information extraction (Finkel et al., 2005), summarization (Aone et al., 1997), question answering (Kafle and Kanan, 2017), just to name a few.

Current research in computer vision and deep learning have emerged great progress in many computer vision problems, especially image classification and image segmentation on the rise of Convolutional Neural Network (CNNs) given enough data. Similar to CNN for computer vision, deep learning (also known as deep structured learning or hierarchical learning) architectures, for examples, Recurrent Neural Networks (RNNs), Long Short-term Memory (LSTM) models and memory-based models, have produced a very high performance (even superior to experts in some circumstances) in vast amounts of NLP tasks.

VQA is a critical and engaging task because it combines two significant fields of computer science: machine vision and NLP. In a VQA task (refer to Figure 1), the inputs are a raw image with no other information and a question about the visual contents of the corresponding image. The goal is to find a short answer to the question (typically a few words or a short phrase).

In a VQA task, machine vision techniques are employed to understand the image, while NLP techniques are to understand the question. In fact, all of the components of a VQA model must be effectively combined to produce the right answer concerning the context of the image. However, this is a challenging task as each component (in a total of four, we will discuss in detail in section 2) in a VQA model must perform well before they are combined. Besides, machine vision and NLP have developed distinct architectures for their tasks.

In VQA, questions can be arbitrary and they may contain many sub-problems in computer vision, for example:

- Object recognition - What is in the basket?
- Object detection - Are there any birds in the water?

– Attribute classification -What color is the surfboard?
– Scene classification - Is this in America?
– Counting - How many kids have their hands up in the air?

Variants include binary (yes/no), multiple-choice and short answer settings. Beyond these, many more complex questions can be asked, for example, the spatial information (what is behind the table?), the common-sense reasoning (why is the boy laughing?), and encyclopedic knowledge about a specific element from the image (how is the egg prepared?). In this respect, VQA establishes a complete AI-task, as it requires multimodal knowledge beyond a single subdomain as well as comforts the increased interest in VQA (Wu et al., 2017).

There are many potential applications for VQA. The most immediate is a computer-human interaction tool for blind people and visually-impaired users as a natural way to query for visual content. A VQA system can also refer to many machine vision sub-tasks such as object recognition, image segmentation, image captioning, decision making to name a few. According to Kafle and Kanan (2017), another obvious application is to integrate VQA into image retrieval systems. This could have a huge impact on social media or e-commerce. VQA can also be used with educational or recreational purposes.

1.2. Motivation

DR or diabetic eye ailment is the primary source of visual impairment in adults aged 20-74 years. It is a condition in which harm jumps out at the blood vessels of the light-delicate tissue at the back of the retina because of diabetes mellitus. The more we have diabetes, and the less controlled our glucose is, the more probable we are to build up this eye entanglement. DR can create in anyone who has type-1 or type-2 diabetes.

Retina recognizes light and changes over it to signals sent through the optic nerve to the brain. When high glucose levels harm blood vessels in the retina, individuals with diabetes can have diabetic retinopathy. In some cases, new irregular blood vessels multiply on the surface of the retina (anomalous fresh recruits vessels develop on the retina), which can prompt scarring and cell misfortune in the retina. Blood vessels can close, preventing blood from going through; or they can swell, release liquid, or hemorrhage. These progressions can take our vision.

According to Solomon et al. (2017), DR positioned as the fifth most regular reason for direct to severe visual debilitation and the fifth most fundamental reason for preventable visual impairment from 1990 - 2010. In 2010, more than 33% of around 285 million individuals worldwide with diabetes had indications of DR, and 33% of these were burdened with vision-debilitating DR. These evaluations were expected to rise further because of the maturing of the populace, expanding predominance of diabetes, and expanding of life expectancy of those with diabetes.

Longer diabetes term and poorer glycemic and circulatory strain control are emphatically connected with DR. It is the primary source of visual deficiency in the working-age population of the developed world. It is estimated to influence around 93 million individuals, 17 million with Proliferative Diabetic Retinopathy (PDR), 21 million with diabetic macular edema, and 28 million with vision-threatening DR around the world. This information features the significant overall general wellbeing weight of diabetic retinopathy and the significance of modifiable hazard factors in its event.

Lee et al. (2015) investigated that progression to vision weakness is challenging to be impeded or averted because DR frequently indicates few symptoms until it is too late to give excellent treatment. Currently, detecting DR is a time-consuming and manual process that requires a prepared clinician to look at and assess digital shading fundus photos of the retina. When human readers submit their reviews, regularly a day or two later, the postponed comes promptly lost development, miscommunication, and prolonged treatment.

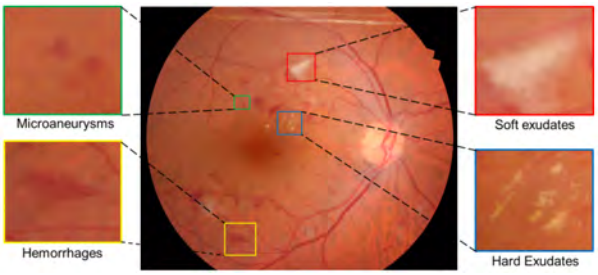


Figure 2: Diabetic Retinopathy: four severe retinal lesions including MA, HE, EX and SE

Figure 2 illustrates key terms for DR. Microaneurysms (MA) are the most constant clinically noticeable changes of diabetic retinopathy. They are localized capillary dilatations which are typically saccular.

Retinal Hemorrhages (HE) is a turmoil of the eye in which bleeding happens into the light touchy tissue on the back mass of the eye. Retinal HE can happen because of shaking, especially in youthful babies (shaken infant disorder) or from severe hits to the head. Hypertension can cause a retinal MA, retinal vein impediment (a blockage of a retinal vein), or diabetes mellitus (which makes little delicate veins shape which is effortlessly harmed.

Cotton-wool spots (CWS), alluded to as Soft Exudates (SE) are nerve fiber layer infarcts or pre-narrow

blood vessel impediments. They are an ischemic occasion of a little measure of tissue. Hard Exudates (EX) represent the collection of lipid in or under the retina optional to vascular spillage. The watery part of the transudative or exudative liquid is retained substantially more quickly than the lipid segment. In this way, the lipid develops in or under the retina and winds up evident as discrete yellowish deposits.

Capillary occlusion happens in both diabetes mellitus and hypertension (HTN); therefore CWS happen in the two conditions. (Hard) Exudate is very typical in diabetic retinopathy (really diabetic macular edema) because of spillage from harmed veins, and VEGF initiated spillage. Exudate is not exceptionally regular in HTN; it is generally just observed in threatening hypertension.

1.3. Objectives

To differentiate between the VQA for natural images and the VQA for retinal images, from this point, I name them "Natural Visual Question Answering" and "Retinal Visual Question Answering", respectively.

This thesis encompasses the following objectives:

- Explore four components of a VQA model, and propose an effective system for the Retinal VQA,
- Propose a fusion scheme for the natural VQA,
- Evaluate the importance of an attention scheme in a VQA model,
- Prepare an image model for the retinal VQA,
- Prepare a question model for the retinal VQA,
- Prepare question/answer groundtruth for the retinal VQA,
- Propose the first and a complete Retinal VQA model for DRS.

1.4. Structure of the Document

The structure of the rest of the document is as follows:

Section 2 presents the Natural VQA including datasets, evaluation metrics, VQA algorithms, and four core components of a VQA model.

Section 3 addresses the first Retinal VQA model including: datasets, groundtruth generation, evaluation metrics, and our proposed model.

Section 4 presents our implementations for the Retinal VQA algorithm including tools used, technical details and image/question/fusion scheme model training.

Section 5 and 6 highlights the results and the discussion, respectively.

Section 7 gives conclusions and future directions for the completion and extension of this work.

2. State of the art

This section presents the Natural VQA including datasets, evaluation metrics, VQA algorithms, and four core components of a VQA algorithm.

2.1. Datasets

2.1.1. COCO-QA

QA sets are made for photos using a figuring that gets them from the COCO dataset. COCO-QA comprises 38,948 testing and 78,736 preparing QA sets. Most questions get some data about the dissent in the photo (69.84%), with a different request being tied in with shading (16.59%), including (7.47%) and region (6.10%). The more critical piece of the request has a single word answer, and there are only 435 unique answers. These restrictions on the proper reactions make evaluation immediate.

The shortcoming of COCO-QA is that it merely has four kinds of request, and these are limited to the kinds of things depicted in COCO's subtitles. The most noteworthy lack of COCO-QA is relied upon to blemishes in the count that was used to create the QA sets. Longer sentences are separated into little pieces for effortlessness of controlling; nevertheless, in tremendous quantities of these cases, the computation does not adjust well to the closeness of explanations and syntactic assortments in sentence progression. These results in cumbersomely communicated request, with various containing semantic faults, and others being jumbled.

2.1.2. The VQA Dataset

The VQA dataset is the most popular dataset for the VQA undertaking. This dataset was discharged as a component of the VQA challenge. It is split into two parts: one dataset includes conceptual clip-art scenes made from models of creatures and people to evacuate the need to process noisy pictures and perform high-level reasoning, and another dataset comprises right pictures from MS-COCO. Inquiries and answers are created from swarm sourced specialists and 10 answers are acquired for each inquiry from unique specialists. Answers are ordinarily a word or a short expression. Roughly 40% of the inquiries have a yes or no answer. For assessment, both multiple choice formats and also open-ended answer generation are accessible.

The first VQA dataset has 204,721 MS-COCO pictures with 614,163 inquiries and 50,000 abstract pictures with 150,000 inquiries. The 2017 cycle of the VQA challenge has a greater dataset with an aggregate of 265,016 MS-COCO and abstract pictures and an average of 5.4 inquiries for each picture. The correct number of inquiries is not said on the challenge site. The VQA daataset comprises both abstract animation images and real images from COCO. Most work on this

dataset has concentrated exclusively on the segment including real-world imagery from COCO, which we allude to as COCO-VQA.

COCO-VQA comprises of three inquiries for each picture, with ten answers for each inquiry. Amazon Mechanical Turk (AMT) laborers were utilized to create inquiries for each picture by being asked to "Stump an intelligent robot," and a different pool of specialists was hired to create the responses to the inquiries. Contrasted with other VQA datasets, COCO-VQA comprises a moderately substantial number of inquiries (614,163 total, with 248,349 for training, 121,152 for validation, and 244,302 for testing). Ten independent annotators then reply each of the inquiries. The different answers per question are utilized as a part of the consensus-based assessment metric for the dataset.

2.1.3. Visual Genome

Visual Genome comprises of 108,249 pictures that arise in both YFCC100M and COCO pictures. It comprises 1.7 million QA sets for pictures, with an average of 17 QA sets for each picture. Visual Genome is the most significant VQA dataset. Since it was just as of late presented, no techniques have been assessed on it past the baselines set up by the creators. Visual Genome comprises of six sorts of 'W' questions: Who, What, When, How, Where, and Why. Two particular methods of information accumulation were utilized to make the dataset.

Visual Genome has significantly more noteworthy answer variety contrasted with different datasets. The 1000 answers that arise most regularly in Visual Genome cover 65% of all answers in the dataset, while they cover 100% for DAQUAR and COCO-QA and 82% for COCO-VQA. Visual Genome's long-tailed allocation is additionally seen in the length of the appropriate responses. Just 57% of answers are single words, compared to 100% of answers in COCO-QA, 90% of answers in DAQUAR, and 88% of answers in COCO-VQA. This variety of answers makes open-ended assessment significantly additionally tricky. Furthermore, because the classes themselves are required to entirely have a place with one of the six "W" types, the assorted variety in the answer may at time artificially stem just from varieties in stating which could be wiped out by provoking the annotators to pick more brief answers. For instance, Where is the motorbike parked? can be replied with "on the street" or all the more compactly with "street."

Visual Genome has no binary (yes/no) questions. The dataset makers contend that this will boost utilizing more complicated questions. This is opposite to The VQA Dataset, where "yes" and "no" are the more regular answers in the dataset.

2.2. Evaluation metrics for Natural VQA

VQA represented as either a multiple-choice or an open-ended task for evaluation. Multiple-choice refers to a task where a VQA model selects an answer from many options, while the output can be a word or a short phrase for the open-ended task. For both tasks, simple accuracy is as follows:

$$\text{simple accuracy} = \frac{\text{correct answers}}{\text{number of questions}} \tag{1}$$

However, the simple accuracy is 'strict' in semantic scenarios as there are wrong answers which might be more acceptable than the others. For instance, a system is asked 'which color of the umbrella in the picture?'. If the correct answer is 'yellow,' then two outputs 'orange' and 'cat' will be penalized in the same way. Due to this problem, some alternatives were proposed for an open-ended task which is robust to inter-human variability in phrasing the answers. One of those is consensus metric (open-ended accuracy). For the VQA dataset, open-ended accuracy is computed as below:

$$\text{open-ended accuracy} = \min(\frac{N}{3}, 1) \tag{2}$$

where N is the number of humans that said the same answer as the algorithm. That is an answer will get the full mark if there are more than three humans gave the same answer. Due to human-agreement, the maximum open-ended accuracy is only 83.3% for the COCO-VQA.

2.3. VQA algorithms

These systems differ significantly in how they integrate the question and image features. Kafle and Kanan (2017) used a survey to assess the various VQA techniques:

- Combining the image and question features using simple mechanisms, e.g., concatenation, element-wise multiplication, or element-wise addition, and then giving them to a linear classifier or a neural network,
- Combining the image and question features using bilinear pooling or related schemes in a neural network framework,
- Having a classifier that uses the question features to compute spatial attention maps for the visual features or that adaptively scales local features based on their relative importance,
- Using Bayesian models that exploit the underlying relationships between question-image-answer feature distributions, and
- Using the question to break the VQA task into a series of sub-problems.

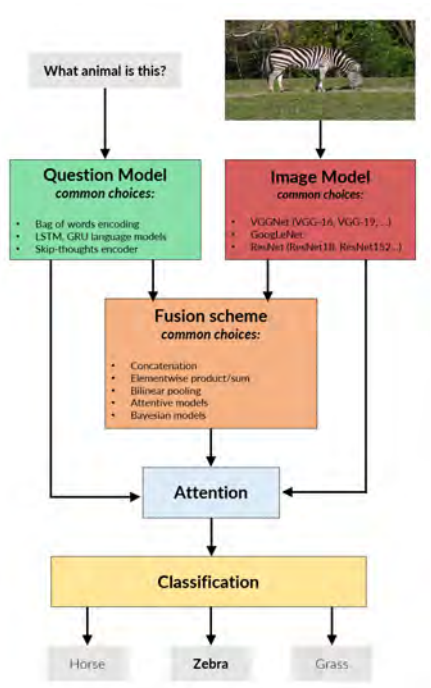


Figure 3: Simplified illustration of the classification based framework for VQA. In this framework, image and question features are extracted, and then they are combined so that a classifier can predict the answer. A variety of feature extraction methods and algorithms for combining these features have been proposed, and some of the more common approaches. are listed in their respective blocks in the figure

To the best of our knowledge, the VQA model includes four learnable elements (refer to Figure 3):

- A question model with the purpose of encoding questions,
- An image model to extract visual feature,
- A fusion model to combine visual features and encoded question,
- An attention scheme to "pay more attention" to specific regions of the given image.

Next, we will discuss each component in detail.

2.4. Question model

The purpose of a question featurization in a VQA method is to encode the input question. A wide options have been developed including bag-of-words (BOW), LSTM encoders (Hochreiter and Schmidhuber, 1997), Gated Recurrent Units (GRU) (Cho et al., 2014), and skip-thought vectors (Kiros et al., 2015).

2.4.1. Bag-of-words

The BOW model is a simplified presentation employed in NLP. In this technique, a sentence or a document is represented by its owns words taking no account of grammar and even word order with the purpose of maintaining multiplicity. For example, given two sentences as follows:

- Henry likes to play soccer. Pete likes soccer too.
- Henry also likes to play volleyball.

Based on these given sentences, BOW algorithm breaks them into two lists of words as follows:

- 'Henry', 'likes', 'to', 'play', 'soccer', 'Pete', 'too'
- 'Henry', 'also', 'likes', 'to', 'play', 'volleyball'

Finally, BOW represents each sentence as a dictionary format:

- {'Henry':1, 'likes':2, 'to':1, 'play':1, 'soccer':2, 'Pete':1, 'too':1}
- {'Henry':1, 'also':1, 'likes':1, 'to':1, 'play':1, 'volleyball':1}

Zhou et al. (2015) adopted BOW in their simple baseline for VQA with no more than 10 lines of codes in Torch. In their work, The input question, the question is first encoded into a vector by the one-hot encoder. Then, the encoded vector is transformed to word vector before concatenating them with the visual vector.

2.4.2. Long Short-term Memory

LSTM blocks were first introduced by Hochreiter and Schmidhuber (1997). They are usually employed to build RNNs' layer. A LSTM has been used successfully in the classification tasks, especially effective for time series tasks such as forecasting and NLP. A common LSTM unit includes four components: a cell, an input gate, an output gate and a forget gate (see figure 4 [1]). In a common LSTM block, the cell is for remembering the signal for a short period; the other gates can be seen as feedforward nodes in a neural network and are connected to the cell.

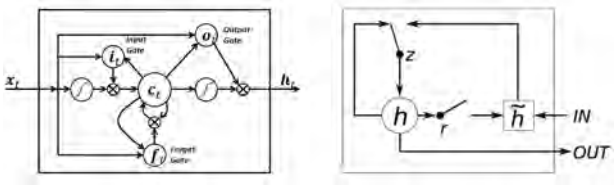


Figure 4: **Left**: a common LSTM block with forget gate includes four components: a cell, an input gate, an output gate and a forget gate. **Right**: a GRU block comprises a cell, an update gate and a reset gate

To summarize, the output of each node can be derived by the set of equations below:

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \quad (3)$$
$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \quad (4)$$
$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \quad (5)$$
$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \quad (6)$$
$$h_t = o_t \circ \sigma_h(c_t) \quad (7)$$

[1] Figure 4 is contributed by user BiObserver under the CC BY-SA 4.0 license at `https://commons.wikimedia.org/wiki/File:Long_Short_Term_Memory.png`

where the initial values are $c_0 = 0$ and $h_0 = 0$ and the operator $\circ$ denotes the Hadamard product (entry-wise product). The subscripts $_t$ refer to the time step.

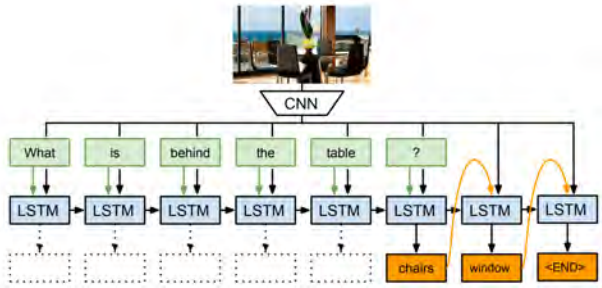


Figure 5: Malinowski et al. (2015) used LSTMs into which each word in the question is fed. The predicted answers (in orange color) were generated by subsequent time-steps.

A number of VQA algorithms have utilized LSTMs to encode questions. Antol et al. (2015) used LSTM to extract word feature from the one-hot encoded question. In Malinowski et al. (2015) (see figure 5), LSTMs were used to be fed each word in the question. Then, the predicted answers (in orange color) were generated by subsequent time-steps. Another example can be found in Gao et al. (2015) where two LSTMs were used: one was fed with visual features, and another one was used to predict the answer.

2.4.3. Gated Recurrent Units

Similar to LSTM, GRU is a gating mechanism for building layer in RNNs. It was developed by Cho et al. (2014). A common GRU has three elements: a cell and two gates (update gate and reset gate), compared to three gates in LSTM (see figure 4). Its performance on time-series tasks was proved to be on par with LSTM, though with fewer parameters

The output can be derived as follows:

$$z_t = \sigma_g(W_z x_t + U_z h_{t-1} + b_z) \tag{8}$$

$$r_t = \sigma_g(W_r x_t + U_r h_{t-1} + b_r) \tag{9}$$

$$h_t = (1 - z_t) \circ h_{t-1} + z_t \circ \sigma_h(W_h x_t + U_h(r_t \circ h_{t-1})) \tag{10}$$

where x_t is input vector, h_t is output vector, z_t is update gate vector, r_t represents reset gate vector and W, U and b: parameter matrices and vector. The operator $\circ$ denotes the Hadamard product. At first, $t = 0$ and the output vector is $h_0 = 0$.

Lu et al. (2017) proposed to use GRU to extract word feature as follows: first, the question is one-hot encoded. Then, they applied a linear transformation to extract the feature vector for each word and sequentially fed into GRU for the last step encoding. At each time-step, the GRU block updates the update and reset gate and outputs a hidden state.

2.4.4. Skip-thought vectors

Skip-though vectors algorithm was introduced by Kiros et al. (2015) and has been used in most of the state-of-the-art VQA models (we will discuss in details in Section 2.6). In this thesis, we also used skip-thought vectors to extract word features.

Skip-though vectors (see Figure 6) is an unsupervised-learning approach for generic, distributed sentence encoder. In this approach, the authors developed an encoder-decoder methodology using the semantic and continuity of text from books with the purpose of reconstructing surrounding sentences given an encoded context. As a result, sentences, which are semantic-related, are represented by a similar vector. During the training, Pearson correlation is incorporated for early stopping.

One of the best advantages of skip-though vectors is the ability to expand vocabulary. In specific, a pre-trained model can be 'fine-tuned' such that our vocabulary can reach up to million words. This feature is crucial for this thesis as we have a huge number of 'unknown' words when we propose to use VQA for the DRS problems (we will discuss how we expand the vocabulary for our project in section 4.3).

Kiros et al. (2015) evaluated skip-thought vectors on eight different tasks: semantic relatedness, paraphrase detection, image-sentence ranking, question-type classification and four benchmark sentiment and subjectivity datasets. In summary, skip-thought vectors can produce highly generic sentence representations that are robust and perform well in practice.

2.5. Image model

An image extractor in a VQA model is responsible for extracting the visual features of the input image. For image features, most algorithms use CNNs that are pre-trained on ImageNet (Krizhevsky et al., 2012). An overview of different VQA approaches that were evaluated on COCO-VQA dataset, and their designs can be seen in Table 1. Table 1 shows that the popular choices for an image model are VGGNet (Simonyan and Zisserman, 2014), ResNet (He et al., 2016), and GoogLeNet (Szegedy et al., 2015).

These common choices for an image model come from that fact that VGGNet, ResNet, and GoogLeNet outperformed the rest on ImageNet challenge. Figure 7 compares the most popular networks with their corresponding top-1 accuracy and number of parameters.

2.5.1. VGGNet

VGGNet was introduced by Simonyan and Zisserman (2014). At the ILSVRC 2014 competition, VGGNet was the runner-up. VGGNet comprises 16-19 (VGG-16 or VGG-19) convolutional layers with an enormous amount of 3×3 fixed size convolution filters which is

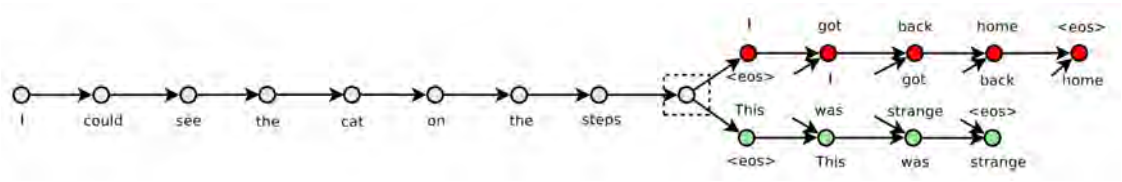


Figure 6: Kiros et al. (2015) trained skip-thought vectors on a huge amount of novels with a diversity genres. In this figure, unattached arrows are connected to the encoder output. Colors indicate which components share parameters. < *eos* > is the end of sentence token.

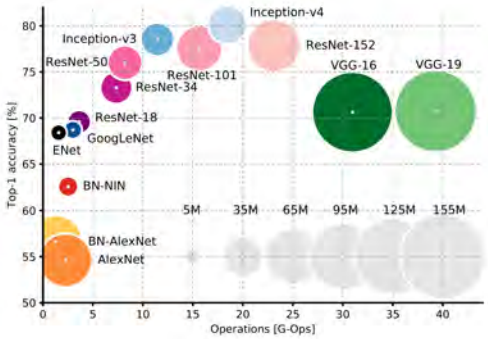


Figure 7: Top-1 accuracy and size of parameters of popular networks on ImageNet challenge. As can be seen, VGGNet, GoogLeNet/Inception and ResNet surpass all other architectures. This figure was prepared by Kafle and Kanan (2017)

Table 1: Overview of different VQA approaches that were evaluated on COCO-VQA and their designs. This table was redistributed from Kafle and Kanan (2017)

Method	Visual	Word	Attention
iBOWIMG	GoogLeNet	BOW	-
SMem	GoogLeNet	BOW	√
SAN	GoogLeNet	LSTM	√
LSTM Q+I	VGGNet	LSTM	-
DPPNet	VGGNet	GRU	-
HieCoAtten	VGGNet	LSTM	√
MCB	ResNet	LSTM	-
MLB	ResNet	Skip-thought	√
MUTAN	ResNet	Skip-thought	√
Proposed	ResNet	Skip-thought	√

similar to AlexNet. Despite its widespread use for image feature extraction, VGGNet is by far the most expensive architecture as it consists of 138 million parameters.

2.5.2. GoogLeNet

Szegedy et al. (2015) won the ILSVRC 2014 with GoogLeNet (Inception) at 6.67% top-5 accuracy which is near human-level performance. LeNet inspired GoogLeNet with the use of batch normalization, image distortions, and RMSprop. This network was diligently designed to have the ability to grow in the depth while maintaining a constant computational

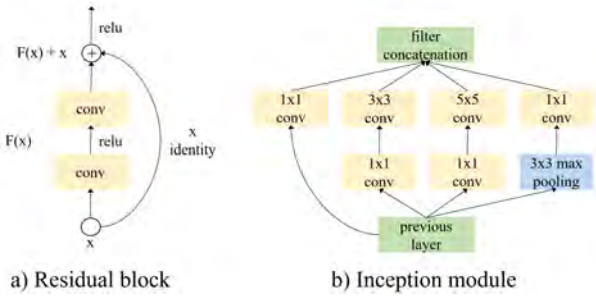


Figure 8: a) He et al. (2016) introduced residual block used in ResNet. b) Inception module was used in GoogLeNet (Szegedy et al., 2015)

complexity. Consequently, the number of parameters dropped from 60 million (AlexNet) to 4 million in a 22-layer network. Their design was inspired by Hebbian principle and the and the intuition of multi-scale processing. In the end, an efficient 'Inception' module (see figure (see figure 8b) was proposed.

2.5.3. ResNet

At the ILSVRC 2015, ResNet, which was developed by He et al. (2016), was the winner as it achieved 3.57% top-5 accuracy that even surpasses human-level performance on this dataset. In ResNet, skip-connection block (see figure 8a) was introduced with the purpose of learning the reference to input layers instead of from unreferenced functions. With the use of these skip-connection blocks, ResNet152, which has a depth of 152 layers, still has fewer parameters than VGGNet, though better performance.

2.6. Fusion scheme

For the task of VQA, the goal is to predict the most likely answer given a question $q \in \mathcal{Q}$ and image $v \in \mathcal{V}$. The problem can be formulated as following:

$$\hat{a} = \underset{a \in \mathcal{A}}{\arg\min}\, p_\Theta(a|\mathrm{q}, \mathrm{v}) \tag{11}$$

where $\hat{a}$ denotes the predicted answer, and Θ represents the whole set of parameters of the model.

The problem solving of learning a classifier, as shown in Equation 11, becomes more straightforward as multimodal pooling, $\Omega(q, v)$, is capable of efficient encoding the relationship between image and question features.

One of the most well-known tools for fusing multimodal features are multimodal pooling. Next, we will explore the state-of-the-art algorithms for the Natural VQA inspired by this technique.

2.6.1. Multimodal Compact Bilinear

Fukui et al. (2016) developed Multimodal Compact Bilinear (MCB) based on the concept of bilinear pooling by applying a linear transformation for every pair of visual and textual features:

$$f_i = \sum_{j=1}^{N}\sum_{k=1}^{M} w_{ijk} q_j v_k + b_i = q^T \mathbf{W}_i v + b_i \tag{12}$$

where q and v are textual and visual features; $\mathbf{W}$ and b_i denote a weight matrix and bias vector for the output f_i, respectively.

Note that the number of parameters for a classifier of Equation 12 is $L \times N \times M$, where L is the number of output features. For example, if $L = N = M = 2000$, the number of parameters will be around 10^{10} which is extraordinary expensive [2].

Thus, MCB, which is inspired by Pirsiavash et al. (2009), proposed the low-rank bilinear mechanism to overcome the parameter explosion issue as below.

As suggested by Pirsiavash et al. (2009), weight matrix can be estimated as $\mathbf{W}_i = \mathbf{Q}_i \mathbf{V}_i^T$ where $\mathbf{Q} \in \mathcal{R}^{N\times d}$ $\mathbf{V} \in \mathcal{R}^{M\times d}$. Here, d is the restriction parameter on the rank of $\mathbf{W}$ which is the key element to compact the rank of output f_i, thus the name 'Multimodal Compact Bilinear'.

Substituting $\mathbf{Q}_i$ and $\mathbf{V}_i$ into Equation 12, we obtain:

$$f_i = q^T \mathbf{Q}_i \mathbf{V}_i^T v + b_i \tag{13}$$

$$= \mathbb{1}(\mathbf{Q}_i^T q \circ \mathbf{V}_i^T v) + b_i \tag{14}$$

where $\circ$ denotes Hadamard product. Next, we redefine $\mathbf{Q} \in \mathcal{R}^{N\times d}$ and $\mathbf{V} \in \mathcal{R}^{M\times d}$ after replacing $\mathbb{1}$ with $\mathbf{P} \in \mathcal{R}^{d\times c}$ to get the projected output f of MCB model as below:

$$\mathrm{f} = \mathbf{P}^T(\mathbf{Q}^T \mathrm{q} \circ \mathbf{V}^T \mathrm{v}) + \mathrm{b} \tag{15}$$

where d and c are hyper-parameters of joint embeddings and the dimension of output, respectively.

2.6.2. Multimodal Low-rank Bilinear

Kim et al. (2016) claimed that bilinear models like MCB, though provide rich representation, tend to be high-dimensional such that limiting the applicability to computationally complex tasks. Hence, Kim et al. (2016) proposed Multimodal Low-rank Bilinear (MLB) algorithm with the purpose of one step further reducing the rank of bilinear by utilizing the concept of bilinear pooling and Hadamard product. MLB outperforms MCB algorithm on all tasks of the VQA dataset. The idea of MLB is as below:

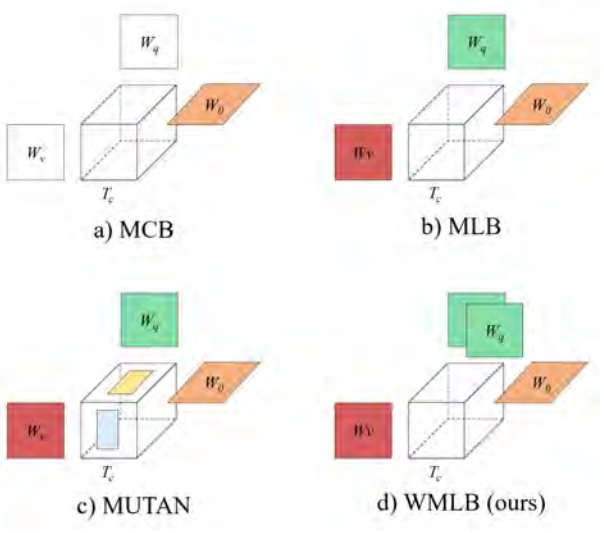


Figure 9: a) MCB: only $\mathbf{W}_o$ is trainable. b) MLB: $\mathbf{W}_q, \mathbf{W}_v$ and $\mathbf{W}_o$ are learnt, while $\mathcal{T}_c$ is fixed. c) MUTAN: all four elements are trained. d) Proposed: similar to MLB with addition tensor $\mathbf{W}_q$

Continuing from Equation 15, the authors suggested that $\mathbf{Q}$ and $\mathbf{V}$ can have their own bias vectors. Hence, Equation 15 becomes:

$$\mathrm{f} = \mathbf{P}^T((\mathbf{Q}^T \mathrm{q} + \mathrm{b}_q) \circ (\mathbf{V}^T \mathrm{v} + \mathrm{b}_v)) + \mathrm{b} \tag{16}$$

It can be rewritten as:

$$\mathrm{f} = \mathbf{P}^T(\mathbf{Q}^T \mathrm{q} \circ \mathbf{V}^T \mathrm{v} + \mathbf{Q}'^T \mathrm{q} \circ \mathbf{V}'^T \mathrm{v}) + \mathrm{b}' \tag{17}$$

where $\mathbf{Q}'^T = \mathrm{diag}(b_v) \cdot \mathbf{Q}^T$, $\mathbf{V}'^T = \mathrm{diag}(b_q) \cdot \mathbf{V}^T$ and $\mathrm{b}' = \mathrm{b} + \mathbf{P}^T(\mathrm{b}_q \circ \mathrm{b}_v)$.

Next, non-linear activations such as sigmoid or tanh are applied to increase the representative capacity of model. Now, the equation above becomes:

$$\mathrm{f} = \mathbf{P}^T(\sigma(\mathbf{Q}\mathrm{q}) \circ \sigma(\mathbf{V}\mathrm{v})) + \mathrm{b} \tag{18}$$

Finally, Kim et al. (2016) proposed to apply the activation function after the Hadamard product as an alternative choice to remove the double gradient calculations. Hence, the final model is formulated as follows:

$$\mathrm{f} = \mathbf{P}^T \sigma(\mathbf{Q}\mathrm{q} \circ \mathbf{V}\mathrm{v}) + \mathrm{b} \tag{19}$$

2.6.3. MUTAN

As discussed in Section 2.6.2, bilinear models are computational expensive. Ben-younes et al. (2017), hence, aimed to parametrize efficiently bilinear interactions between image and question features. MUTAN was inspired by the concept of multimodal

[2] A model with 10 billion float32 scalars needs 40Go to hold, while a Titan 1080 Ti hold about 12Go

tensor-based Tucker decomposition. To the best of our knowledge, currently, MUTAN is the state-of-the-art algorithm of Natural VQA.

In Tucker-decomposition domain, Equation 12 can be reformed:

$$\mathrm{f} = (\mathcal{T} \times_1 \mathrm{q}) \times_2 \mathrm{v} \tag{20}$$

where $\mathcal{T} \in \mathcal{R}^{d_q \times d_v \times \mathcal{A}}$, operator $\times_i$ denotes the *i-mode* product between a tensor and a vector. Full tensor $\mathcal{T}$ is then Tucker decomposed into 3-way tensor to obtain:

$$\mathcal{T} = ((\mathcal{T}_c \times_1 \mathbf{W}_q) \times_2 \mathbf{W}_v) \times_3 \mathbf{W}_o \tag{21}$$

where $\mathbf{W}_q \in \mathcal{R}^{d_q \times t_q}$, $\mathbf{W}_v \in \mathcal{R}^{d_v \times t_v}$ and $\mathbf{W}_o \in \mathcal{R}^{\mathcal{A} \times t_q}$

In general, we can summarize: $\mathcal{T} = \{\mathcal{T}_c, \mathbf{W}_q, \mathbf{W}_v, \mathbf{W}_o\}$.

Figure 9 compares the tensor designs between MCB, MLB, MUTAN and our proposed algorithm which will be discussed in Section 3.4.

The literature on Ben-younes et al. (2017) has shown that the intra-modal projection matrices, $\mathbf{W}_q^{mcb}, \mathbf{W}_v^{mcb}$ in MCB are fixed diagonal, $\mathcal{T}_c$ is a sparse fixed tensor, and only the $\mathbf{W}_o$ is learnable.

In the analysis of Tucker decomposition, Ben-younes et al. (2017) also found that bilinear interaction, which is used in MLB, corresponds to a canonical decomposition of the tensor where $\mathbf{W}_q^{mcb} = \{\mathcal{I}_R, \mathbf{W}_q, \mathbf{W}_v, \mathbf{W}_o\}$ and $t_q = t_v = t_o = R$. That is three elements $\mathbf{W}_q, \mathbf{W}_v$ and $\mathbf{W}_o$ are learnt, while $\mathcal{T}_c$ is fixed.

Different from MCB and MLB, MUTAN was developed such that all of four components of tensor $\mathcal{T} = \{\mathcal{T}_c, \mathbf{W}_q, \mathbf{W}_v, \mathbf{W}_o\}$ are trainable, while $\mathcal{T}_c$ is decomposed by low-rank Tucker.

2.7. Attention scheme

Explored by Fukui et al. (2016), attention scheme utilize a distribution probability α over $S \times S$ lattice place. Attended visual feature is then defined as:

$$\hat{v} = \|_{g=1}^{G} \sum_{s=1}^{S} \alpha_{g,s} \mathbf{F}_s \tag{22}$$

where $\|$ denotes the concatenation of G glimpses, $\alpha_{g,s}$ represents the distribution probability of glimpse g on s region.

Notice that image features extracted from the last convolution layer of ResNet152 have the dimension of $14 \times 14 \times 2{,}048$. Fukui et al. (2016) proposed to use two convolutional layers to predict the attention weight for each 14×14 grid. Then, the authors took a weighted sum of the spatial vectors with the normalized soft attention map to generate the attended visual representation. Finally, they suggested employing multiple "glimpses" before being merged with the language representation to produce sufficient attention to salient locations.

In this thesis, we also use attention mechanism to obtain a better-fused representation. Figure 10 demonstrates a scheme to generate attention heatmap.

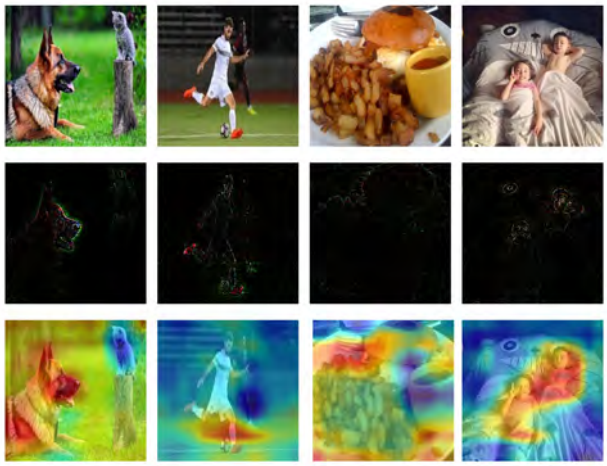


Figure 10: The original image is shown on the top. The middle image represents guided Grad-Cam. The bottom image shows heatmap for attention scheme.

3. Retinal Visual Question Answering

This section addresses our work - the first Retinal VQA. This section comprises four sub-sections: datasets, VQ pair groundtruth generation, evaluation metrics and our proposed method. The detail of implementation will be discussed later in Section 4.

3.1. Datasets

3.1.1. Kaggle - Diabetic Retinopathy Detection

DR is the most common reason for avoidable vision weakness, mostly influencing working age population. Late research has given a superior comprehension of necessity in clinical eye mind practice to recognize better and less expensive methods for identification, administration, determination, and treatment of retinal sickness. The significance of DR screening projects and trouble in accomplishing the dependable early determination of diabetic retinopathy at a sensible cost needs thoughtfulness regarding create computer-aided diagnosis tool.

Notice the danger of DR, California Healthcare Foundation sponsored the challenge Kaggle DR Detection in 2015 with prestigious $100,000 prize pool. The purpose of this challenge is to classify DR images into five possible grades from 0 to 4. Submissions were evaluated by the quadratic weighted kappa metric, which measures the agreement between two ratings. Typically, the value of the weighted kappa lies in the range from -1 to 1, where -1 and 1 represent the complete disagreement and agreement between two raters.

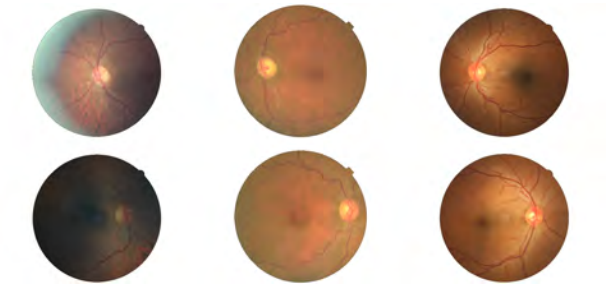


Figure 11: Three sample pairs from Kaggle DR Detection. **Top**: left field. **Bottom**: right field.

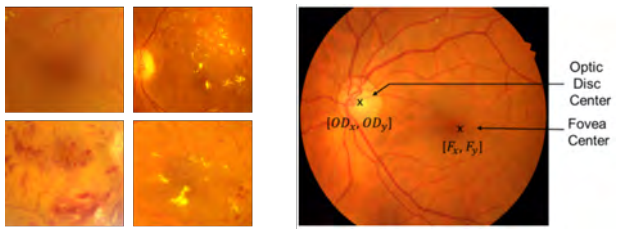


Figure 12: **Left**: from left to right and top to bottom: (a) Grade 1: Mild - non-proliferative diabetic retinopathy (NPDR), (b) Grade 2: Moderate - NPDR, (c) Grade 3: severe - NPDR, and (d) Grade 4: proliferative diabetic retinopathy (PDR). **Right**: an example of optic disc and fovea coordinates

The challenge organizers provided a large, high-resolution dataset. In particular, the dataset comprises around 25,000 pairs (left and right) retinal images, where each image has a resolution of $4,752 \times 3,168$. One of the challenges of this competition is that the images came from different modality cameras which affected the consistency of the dataset. Another challenge is the presence of noise and variation within the dataset (refer to Figure 11). Hence, preprocessing dataset plays a vital role in this challenge.

For each image, a clinician rated the severity of DR on a scale of 0 to 4 as follows:

- No DR
- Mild
- Moderate
- Severe
- PDR

In this thesis, we used the dataset of Kaggle DR Detection to build an image model for the Retinal VQA. The details of training will be discussed later in Section 4.4.

3.1.2. Indian Diabetic Retinopathy Image Dataset

Indian Diabetic Retinopathy Image Dataset (IDRID) is a challenge organized by IEEE International Symposium on Biomedical Imaging (ISBI-2018), Washington D.C. The challenge comprised two phases and was divided into three sub-challenges: (i) lesion segmentation, (ii) disease grading and (iii) optic disc and fovea detection. In this challenge, the organizers provided the high-quality dataset with the resolution of $4,288 \times 2,848$.

Lesion segmentation is the first sub-challenge of IDRID. In this sub-challenge, participants were asked to segment images into four severe retinal lesions (four sub-tasks) including MA, HE, EX and SE (see Figure 2). The dataset for this sub-challenge includes two cases: apparent retinopathy (81 images) and no apparent retinopathy (NAR) (89 images). Then, it was divided into two sets: train (67%) and test set (33%).

The second sub-challenge refers to disease grading (see the left of figure 12). Similar to Kaggle DR Detection challenge, this sub-challenge aims to detect the disease severity level of DR and Diabetic Macular Edema (DME). In this sub-challenge, the participants build models to differentiate the grades of DR (five levels) and DME (three levels). The dataset includes 516 images with expert annotations.

The last sub-challenge refers to the optic disc (OD) and fovea detection (see the right of Figure 12). The purpose of this sub-challenge is to localize the OD and fovea coordinates from the retinal images. In this sub-challenge, a total of 516 images were given.

In this thesis, all of the QA pairs were generated from the first sub-challenge dataset (refer to Section 3.3). Besides, we trained two of our image models on the dataset of the second sub-challenge corresponding to two disease severity level of DR and DME. The details of training process will be discussed in Section 4.4. We can also expand the QA-pair groundtruth from the fovea and OD coordinates groundtruth; however, due to time limitation, we will leave this task for future work.

3.2. Evaluation metrics for Retinal VQA

Section 2.2 elaborated on the evaluation metrics used in Natural VQA. In this section, we present the evaluation metrics for Retinal VQA.

Different from popular COCO-VQA with 600,000 questions/answers (QA) for train/valid/test sets in total, in this thesis, we generated the questions/answers-pair groundtruth on our own based on the segmentation groundtruth from IDRID (see Section 3.3 for more details). That explains why open-ended accuracy is not appropriate for retinal task.

Instead, we propose to use two evaluation metrics: simple accuracy (refer to Equation 1) and weighted accuracy.

Section 3.3 indicated that we have a total of 222,360 QA pairs with severe imbalanced yes/no/undefined answer. In specific, most answers are 'no' (83%), with other answers being 'undefined' and 'yes', 11% and 6%, respectively. In this case, we propose to use weighted accuracy as following:

$$\text{weighted accuracy} = \frac{1}{n}\sum_{i=0}^{n}\frac{T_i}{N_i} \tag{23}$$

where T_i denotes the number of correct predictions for the answer i, for example, 'yes', and N_i represents the total number of answer i. Using weighted accuracy, we ensure that 'yes,' 'no' and 'undefined' answers have the same weight.

3.3. QA-pair groundtruth generation

This section outlines our approach generating QA pair groundtruth for the Retinal VQA. For this task, it is crucial to generating a diverse set of QA pairs that can match the DR problem, thus evaluate the feasibility of Retinal VQA. The ability to expand the groundtruth for data-demanding models such as deep learning models and out-of-vocabulary also plays a vital role.

As mentioned in Section 3.1, all of our QA pair groundtruth so far has come from the sub-challenge 2 - lesion segmentation - of IDRID. Before acquiring auto-generate QA pair groundtruth, there is a further step to preprocess and clean the binary groundtruth images before being fed into groundtruth generation program. This step is described below.

Segmented groundtruth images of four cases (microaneurysms, soft exudates, hard exudates, and hemorrhages) are initially transformed to have similar affine transformation as the preprocessed image which is used to train image model. To be clear, we trained our image model before generating QA groundtruth. During the task of image model training, we have applied many preprocessing techniques to obtain a normalized and consistent dataset before feeding into ResNet (see Section 4.4) using ImageMagick which is a low-level command-line image processing tool. That is we have to do reverse engineering to transform groundtruth image to get the same transformation of the normalized image. Figure 13 summaries our reverse engineering technique to transform segmented image to get the similar affine transformation of the processed image.

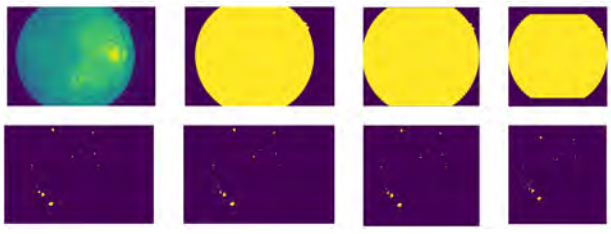


Figure 13: Top: original image, bottom: segmented image. From left to right: original, convert to gray-scale, remove left and right background, pad background to top and bottom borders to get square resolution

After getting transformed groundtruth images, we are ready to generate QA pairs automatically. At this point, we propose to create three types of questions and one type of answer (yes/no/undefined). In particular, three types of questions include (i) Is there X in the fundus? (ii) Is the X larger/smaller than the Y? (iii) Is there any X in Z location? As can be seen, the first type of question represents a classification/segmentation task, while the second and the third kinds describe a semantic/quantitative segmentation and localization task, respectively. In this sense, VQA proves how challenging it is by referring to three sub-tasks of computer vision at the same time. In fact, the types of questions can be expandable, thus becomes more complicated. An example is the COCO-QA dataset of Natural VQA with more than 60 types of questions.

Next, we investigate how each type of questions are generated.

First, "Is there X in the fundus?" is a classification/segmentation question where the possible answers are "yes" or "no." An example of this is "Is there microaneurysms in the fundus?" Generating this type of QA groundtruth is straight-forward: we count the number of non-zero pixels in a binary groundtruth. If this number is greater than 0, we return the answer "yes," if not, we mark as "no."

Second, "Is the hard exudates larger than the microaneurysms?" is an example of the second type of semantic/quantitative segmentation task. For this type of question, we also count the number of non-zero pixels in hard exudates and microaneurysms binary images. If hard exudates have more non-zero pixels than microaneurysms, we mark 'yes'; otherwise, the answer is 'no.' Different from the first kind of question, the second type comprises another possible answer which is "undefined." "Undefined" is set when we examine the NAR image where there is no sign of DR.

Third, "Is there any X in Z location?" is the last type of question that we generate for the Retinal VQA problem which asks the VQA algorithm to be capable of localizing and decoding. One example can be seen in the following sample: "Is there any microaneurysms in 0_0_16_16 location?" Here, 0_0_16_16 is an encoded location where 0_0 is x, y coordinates and 16_16 is the size of the moving window. Again, the groundtruth generation task becomes straightforward as the expected answer is "yes" when there is X in the 16 $times$16 window, and "no," otherwise.

Note that, in total we have 170 images for QA generation (81 DR + 89 NAR). That is with the moving square window size 16; we can generate about 220,000 QA pairs which is sufficient to train a VQA model. However, as mentioned above, the QA groundtruth will be more diverse if there are more types of questions and, of course, more images in the dataset. We leave this task for our future work.

Last but not least, we strictly follow COCO-QA to

```
sample =
{
    "question_id": "0000000000",
    "image_name": "IDRiD_01.jpg",
     "question":  "Is the haemorrhages larger
than the microaneurysms?",
    "answer": "no",
    "answers_occurrence: [["no", 10]]
}
```

Figure 14: A sample of QA groundtruth

obtain standard QA pair groundtruth (refer Figure 14). To elaborate on that, we first adopt Pandas dataframe to construct a table of groundtruth where each row refer to a unique question ID. Next, we reconstruct this dataframe to a JSON format which is a list of dictionaries where the keys and values of a dictionary can be seen above. Here, answer occurrence is set to 10 because the groundtruth is auto-generated but not manual in the Natural VQA.

3.4. Weighted Multimodal Low-rank Bilinear Attention Network

In this section, we introduce a simple yet efficient VQA algorithm which outperforms state-of-the-art bilinear models in both Natural and Retinal VQA datasets. WMLB was inspired by MLB in the sense of low-rank bilinear pooling.

Recall that a full tensor $\mathcal{T}$ can be Tucker decomposed into three-way tensor as below:

$$\mathcal{T} = ((\mathcal{T}_c \times_1 \mathbf{W}_q) \times_2 \mathbf{W}_v) \times_3 \mathbf{W}_o \tag{24}$$

where $\mathbf{W}_q \in \mathcal{R}^{d_q \times t_q}$, $\mathbf{W}_v \in \mathcal{R}^{d_v \times t_v}$ and $\mathbf{W}_o \in \mathcal{R}^{\mathcal{A} \times t_q}$

Kafle and Kanan (2017) pointed out that question only models perform substantially better than image models. Inspired by this study, we propose to use a weighted model such that question have more weight than image features. We have tried with many combinations; finally, the best combination turns out to be that question's weight doubles image's. In short, our model can be formulated:

$$\mathcal{T} = ((\mathcal{T}_c \times_1 (\mathbf{W}_q \circ \mathbf{W}_q)) \times_2 \mathbf{W}_v) \times_3 \mathbf{W}_o \tag{25}$$

Intuitively (refer to Figure 9), we take the Hadamard product of question feature with itself before being Tucker decomposed to train VQA model.

Overall, Figure 15 illustrates our proposed VQA algorithm. Notice that, our fusion scheme is fed with (i) visual features, which are the outputs of the last convolution layer of ResNet152 and have $14 \times 14 \times 2,048$ dimensions, and (ii) word features with the dimension of $2,400$ produced by skip-thought vectors.

We employ the same kind of multi-glimpse attention mechanisms proposed by Fukui et al. (2016). We first employ WMLB to calculate the word embeddings score. Together with visual features, we then compute the global image features with $1 \times 1 \times 2,048$ of dimension by taking the weighted sum of these scores. After that, we fuse global image and question features by using WMLB to output an N-size vector where N represents the number of top answers. Finally, we can predict the answer by a softmax function.

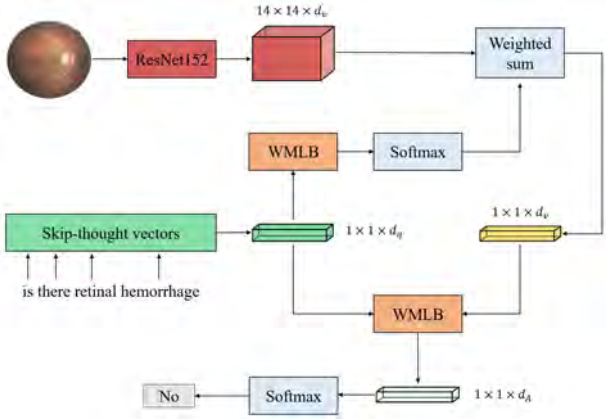


Figure 15: Our proposed WMLB attention scheme for Retinal VQA. This figure was redistributed from Ben-younes et al. (2017)

4. Retinal Visual Question Answering Implementations

4.1. Tools

As most of this thesis was from a distance and we have to handle extensive datasets, operating-system-level virtualization and low-level processing images tool are essential for this project. Our project is developed in Python, and we have used the following libraries and tools:

First, Docker is operating-system-level virtualization and mainly developed for Linux. Docker allows parallel independent containers by isolating the available resources that can package dependencies and application. For example, we can query for more resources (RAM, GPU, and CPU) for more cumbersome tasks. In case of a light task, we can ask Docker for a sufficient resource without stopping the work from the others.

Second, ImageMagick is a low-level image processing tool with the purpose of converting, transforming and editing raw images with high speed, thanks to parallel and multi-threading. ImageMagick can be called by *bash* command in Linux. In this project, we use ImageMagick for preprocessing images.

Third, Augmentor is a Python image augmentation library which allows for more exceptional grained control over augmentation. In this thesis, Augmentor

was responsible for automating image augmentation to balance classes in the dataset.

4.2. Technical details

The proposed method has been implemented in the Python language, using PyTorch. All experiments have been run on a GNU/Linux machine box running Ubuntu 16.04, with 32 GB RAM. CNN training has been carried out on two GTX 1080 Ti (NVIDIA Corp, United States) with 24 GB RAM in total.

4.3. Question model training

Recall that the purpose of a question model in a VQA method is to encode the input question. In this thesis, we use skip-thought vectors to extract word features.

One of the challenges of transferring VQA to the retinal domain is to make question model 'understand' new words which have not been seen before. To overcome this shortcoming, there are two options. The first option is to train skip-thought from the scratch which is time-consuming. Kiros et al. (2015) proposed the second option which was inspired by the "Translation Matrix." In particular, the authors constructed a mapping matrix from the trained Word2Vec, Ω_{w2v}, to a word embedding RNN, Ω_{st}. Notice that the vocabulary in Ω_{w2v} is much greater than in Ω_{st}.

This simple vocabulary expansion method proposed in skip-though vectors is used in this thesis. Next, we will describe in detail how we train the mapping, $\mathbf{q}_{st} = \mathbf{W}\mathbf{q}_{w2v}$, to extend vocabulary to the scope of Retinal VQA.

First, a Word2Vec model trained on Google News covering up to 100 billion words is employed. This pre-trained model contributed by Mikolov et al. (2013) ensures that unseen words in Retinal VQA was learnt.

Second, we train a linear regression model, $\mathbf{W}$, without regularization to map the Word2Vec embedding space linearly above to the skip-thought embedding space.

Finally, we apply the mapping $\mathbf{q}_{st} = \mathbf{W}\mathbf{q}_{w2v}$ to generate word features.

4.4. Image model training

The choice of ResNet152 (see Figure 16) for our model comes from the fact that ResNet produces superior performance over VGGNet or GoogLeNet across multiple algorithms (Kafle and Kanan, 2017). This is evident from the models that use identical setup and only change the image representation.

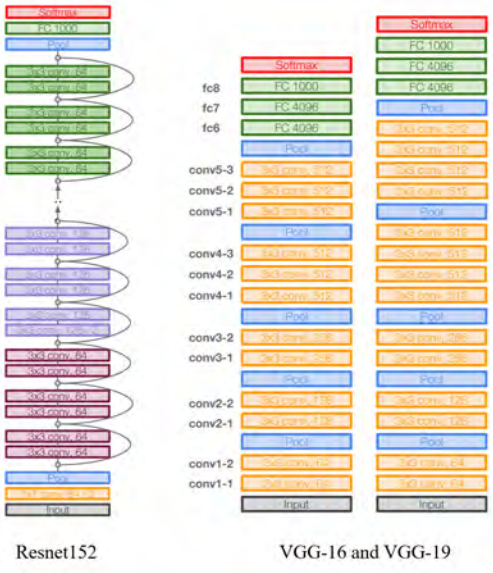


Figure 16: ResNážźt2, VGGNet-16 and VGGNet-19. This figure was redistributed from lecture notes prepared by Fei-Fei Li.

4.4.1. Preprocessing

Both Kaggle DR and IDRID datasets come from different models and types of cameras, thus affect the visual appearance. We encounter noise and variance as some images contain artifacts, be out of focus, underexposed, or overexposed. Consequently, preprocessing techniques, that normalize images into a range that similar or standard to the distribution of the whole dataset, is essential. We adopt low-level language ImageMagick for this task due to large datasets (around 50,000 high-resolution images in total). In the end, we manage to preprocess both datasets in only one hour.

In this thesis, we apply a series of preprocessing techniques to standardize both datasets. These algorithms are called sequentially:

1. **-fuzz 10% -trim +repage**: this command initially searches for a target color (in this case, background) and mark all neighbor pixels within a distance as equivalent. Then, it removes the background borders to obtain an only-object frame.
2. **-extent $size** is used to pad the borders (top/bottom or left/right) with background color to obtain a square frame.
3. **-background black** sets the background color to black.
4. **-gravity center** centers the object within image.
5. **-equalize -colorspace RGB** performs histogram equalization on the image channel-by-channel.
6. **-modulate brightness** (optional) varies the brightness, saturation, and hue of an image.
7. **-sigmoidal-contrast contrastxmid-point** (optional) increases the contrast of an image with-

out saturating highlights or shadows. Here, contrast denotes the level to increase the contrast, and the mid-point represents where the maximum change 'slope' in contrast should fall in.

4.4.2. Balancing dataset and image augmentation

Both Kaggle DR and IDRID datasets contain five DR grades (IDRID also provided three DME levels), and the distribution of classes in both cases was imbalanced (see table 2). This leads to biased and inaccurate supervised model. To cope with this, we propose to use an up-sampling technique, into which image augmentation is incorporated, to enable the balance of training set.

Table 2: Imbalanced classes in Retinal VQA datasets

Level	Kaggle DR	IDRID DR	IDRID DME
0	25810	168	222
1	2443	25	51
2	5292	168	243
3	873	93	
4	708	62	

First, we split the training set into two parts: training (80%) and validation (20%) with corresponding to proportions in each class.

Second, we apply image augmentation through different ways including rotation (10 degrees), flip left/right and flip top/bottom on the classes, which has fewer samples, to acquire balanced dataset. For example of the Kaggle DR (see table 2), we fix class 0 and apply image augmentation on the others to obtain 25810 samples for each class.

4.4.3. Evaluation metric

Inspired by Kaggle DR challenge, the evaluation metric for training model is the quadratic weighted kappa which measures the agreement between actual and predicted labels. Typically, the value of the weighted kappa lies in the range from -1 to 1, where -1 and 1 represent the complete disagreement and agreement between two raters - A (groundtruth) and B (predicted).

First, an $N \times N$ matrix $\mathcal{H}$ is formulated such that $\mathcal{H}_{ij}$ refers to the number of images that were derived a rating i and j by A and B, respectively.

Second, an $N \times N$ matrix of weights w is then calculated:

$$w_{ij} = \frac{(i-j)^2}{(N-1)^2} \tag{26}$$

Third, assume that there is no correlation between rating scores, an $N \times N$ histogram matrix of expected ratings $\mathcal{E}$ is computed as the outer product between each rater's histogram vector of ratings, normalized such that $\mathcal{E}$ and $\mathcal{H}$ have the same sum.

Finally, from these three matrices above, the quadratic weighted kappa is calculated:

$$\kappa = 1 - \frac{\sum_{ij} w_{ij}\mathcal{H}_{ij}}{\sum_{ij} w_{ij}\mathcal{E}_{ij}} \tag{27}$$

4.4.4. Training

In this project, we use Adam optimizer with a learning rate of 10^{-4} and momentum of 0.9. Batch size is set to 512. We also adopt early stopping to cope with overfitting. The loss function is cross entropy, while the evaluation metric is the quadratic weighted kappa.

As the dataset is significant, we train three ResNet152 models (one on KaggleDR and two on IDRID DR and DME) from scratch. The training time for Kaggle dataset is around 11 hours, while it takes 45 minutes for IDRID cases. At the end, the quadratic weighted kappa for training and validation sets are 0.86 and 0.4, respectively.

4.5. Fusion and attention schemes training

After extracting image and question features from ResNet152 and skip-thought vectors, we feed them into the fusion block as shown in Figure 15. This section describes the training process of WMLB with attention mechanism.

4.5.1. Preprocessing

Image

ResNet152 first extracts image features before being stored in HDFStore format where the key name is the image ID. One of the advantages of HDFStore is its dict-like data store such that during training batch of image feature vectors are easily retrieved provided the list of image IDs.

Question

We initially clean QA pair groundtruth. This process includes: make all characters lowercase, remove periods and articles, convert number words to digits, add apostrophe if a contraction is missing, for example, 'dont' becomes 'don't,' replace all punctuation with space.

Next, we tokenize each sentence into a list of words. For example, "Is the retinal hemorrhage larger than the hard exudate?" becomes ['is', 'the', 'retinal', 'hemorrhage', 'larger', 'than', 'the', 'hard', 'exudate'].

Then, we generate the vocabulary from the QA groundtruth.

Finally, we save word features to pickles after extracting them from the skip-thought method.

4.5.2. Loss function

We employ cross entropy loss for multi-class training. It can be seen as belows:

$$L(p, y) = \sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{28}$$

where M denotes the number of classes, y represents the binary indicator (0 or 1) if class label c is the correct classification for observation o, and p is the predicted probability observation o is of class c.

4.5.3. Training

Similar to image model training, we employ Adam optimizer with learning rate and momentum of 10^{-4} and 0.9, respectively. Batch size is set to 512 for no attention and 128 for attention models. The reason is that the later models are very memory-consuming.

We also adopt early stopping to cope with overfitting. The loss function is cross entropy, while the evaluation metric are simple and weighted accuracies. The training time for no attention networks is around 4 hours, while it takes 40 hours for attention architectures.

5. Results

5.1. Natural Visual Question Answering

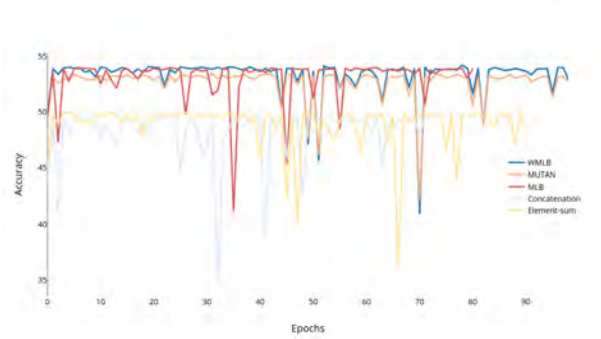


Figure 17: Comparison of performance of five models over training epochs on VQA v1 dataset

Figure 17 and 18 present the top-1 accuracy of VQA v1 and v2 test-set of our model to compare with other models over epochs.

Interestingly, Figure 17 illustrates that the accuracies of all models on VQA v1 tend to converge much faster compared to v2 (Figure 18). Similar behavior can also be seen in the Retinal VQA dataset (refer to Figure 20). This correlation is related to the diversity and the number of questions in the groundtruth.

From the Figure 18, we can see those bilinear models outperform other models by 7%. MUTAN, which is the current state-of-the-art bilinear model comprising four learnable components, tends to fluctuate dramatically before convergence. After going up gradually, our model remains steady from epoch 40 with the accuracy of 51%. Note that the reported metric is top-1 or simple accuracy which is calculated by Equation 1.

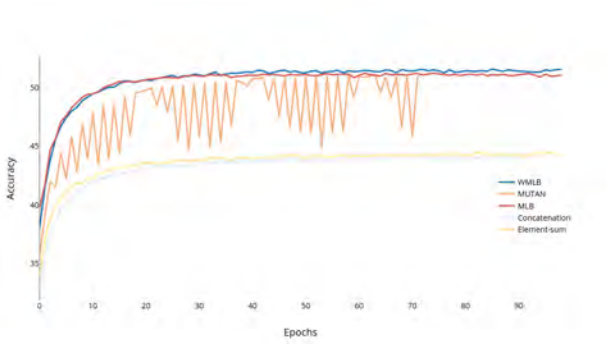


Figure 18: Comparison of performance of five models over training epochs on VQA v2 dataset

Table 3 compare out WMLB with other single models on VQA test-dev on open-ended accuracy. It can be seen from the data in Table 3 that attention mechanism has boosted the performance by around 3%. The overall accuracy of our model is approximately 0.04% and 0.31% above the next best model on the Open-Ended task of VQA v1 and v2, respectively. The major improvements are from yes-or-no (Y/N) on VQA v1 and number (No.) on VQA v2.

5.2. Retinal Visual Question Answering

Figure 19 presents the summary statistics of the training process of our approach WMLB on the Retinal VQA dataset. Similar to VQA v1, the performance on validation is likely to rise sharply before fluctuating around a steady line.

Figure 19 compare our approach with the two state-of-the-art methods. The first observation we can make is that our method WMLB produces superior performance over MLB and MUTAN as it reaches the peak at 91.92%, while the maximum accuracies are 90.97% (MLB) and 90.82% (MUTAN). Interestingly, MLB and our approach WMLB tend to reach the peaks rapidly (at epoch 10 and 30) compared to 95 of MUTAN.

Table 4 compares the simple and weighted accuracies of our model to other bilinear models. What is striking in Table 4 is the the ability to deal with bias problem in our approach (also refer to Figures 20 and 22). In particular, it provides the highest number of correct 'Yes' and 'Undefined' answers in a great imbalanced groundtruth. As a result, WMLB outperforms all the previous methods on the test set in both evaluation metrics: simple and weighted accuracies.

Table 3: Grounding accuracy on VQA v1 and v2 datasets

Model	Params	VQA v1				VQA v2			
		Y/N	No.	Other	All	Y/N	No.	Other	All
Concat	8.9	79.25	36.18	46.69	58.91	-	-	-	-
MCB	32	80.81	35.91	46.43	59.4	-	-	-	-
MLB_noatt	3.7	82.64	34.93	46.49	58.57	77.44	46.34	36.19	56.69
MUTAN_noatt	3.4	47.02	35.37	47.02	58.9	77.37	46.33	36.32	56.68
WMLB_noatt	3.7	81.9	34.6	46.78	58.39	77.02	46.39	35.97	56.54
MLB	7.7	83.87	35.63	51.86	61.81	80.25	51.35	37.87	60.44
MUTAN	4.9	83.16	36.28	51.03	61.2	79.47	51.59	37.62	60.23
WMLB (ours)	7.7	83.96	35.75	51.84	**61.85**	80.14	52.09	37.75	**60.75**

Table 4: Grounding accuracy on Retinal VQA dataset. Note that (*) denotes the number of correct answers, and 'Un.' represents 'Undefined'

Model	Params	Accuracy				Weighted accuracy			
		No*	Un.*	Yes*	All	No	Un.	Yes	All
MLB_noatt	3.7	37015	2996	76	87.56	95.67	59.44	3.71	52.94
MUTAN_noatt	3.4	38246	2996	7	90.10	98.85	59.44	0.34	52.88
WMLB_noatt	3.7	36386	3608	0	87.36	94.05	71.59	0	55.21
MLB	7.7	37018	4480	157	90.97	95.68	88.89	7.65	64.07
MUTAN	4.9	37441	3925	210	90.82	96.77	77.88	10.24	61.63
WMLB (ours)	7.7	37249	4480	355	**91.92**	96.28	88.89	17.31	**67.49**

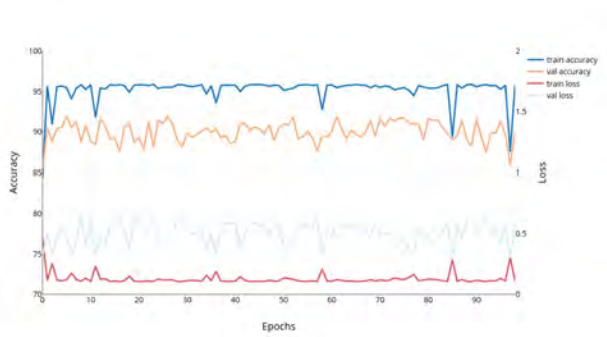


Figure 19: Train/val accuracy and loss over epochs of our method on the Retinal VQA dataset

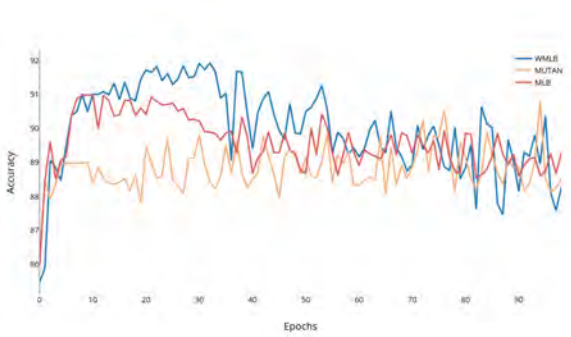


Figure 20: Comparison of performance of thee bilinear models on the validation set on the Retinal VQA dataset

6. Discussion

Despite many proposed methods for VQA, it is hard to argue which general techniques outperform the rest, though bilinear pooling methods are current state-of-the-art. As discussed, a VQA system comprises four learnable components: question model, image model, fusion and attention schemes. To improve the performance, hence, we have four options to refine. In this section, we analyze further four elements in a VQA method and investigate a possibility to reduce training time.

The first element of a VQA system is the visual extractor. Data from several studies suggest that ResNet produces superior performance over VGGNet or GoogLeNet across multiple algorithms. This leads

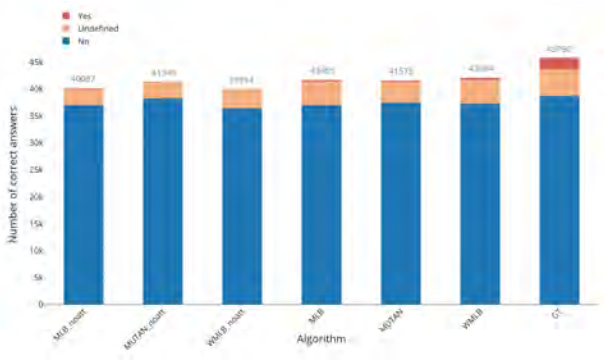


Figure 21: Number of correct answer per type in different models

to a concern: "Is there any visual extractor which provides better performance than ResNet?" The answer is yes. Anderson et al. (2017) published a paper in which they described how image features from bottom-up attention outperform traditional grid-like feature maps from a CNN. Figure 23 demonstrates that Top-down visual attention mechanisms enable deeper image understanding through fine-grained analysis and even multiple steps of reasoning.

The second question in this study sought to determine the importance of the word model which is one of two distinct data streams in VQA. This thesis further supports that a weighted model outperforms a 'balance' model with the same setup. In other words, current VQA systems are more dependent on the question than visual content. A possible explanation for this might be that the Natural VQA datasets tend to have a strong bias on textual content.

Fusion and attention schemes can be combined in one block to be the brain of a VQA system. This thesis confirms that models with attention network are superior than without attention architectures (Table 3 and 4). However, Kafle and Kanan (2017) pointed out that attention alone is not sufficient for a good VQA, but incorporating attention into a strong fusion baseline does.

One of the weaknesses of WMLB, in specific, and bilinear models, in general, is the training time. In particular, it currently takes WMLB 50 hours to train on VQA v2 dataset on 100 epoch. To boost the performance by 1-2%, several studies have revealed that data augmentation from Visual Genome, which triples the size of our training set, would support. That is the training time might take 150 hours for two Titan 1080 Ti GPUs and 300 for a single GPU which is dramatically exhaustive. To overcome this shortcoming, Teney et al. (2017) suggested a list of findings including sigmoid outputs, soft training targets, image features from bottom-up attention, gated tanh activations, output embeddings initialized using GloVe and Google Images, large mini-batches, and smart shuffling of training data. Employing these findings in their study, Teney et al. (2017) won the first place in the 2017 VQA Challenge, though with a relatively simple model and considerable training time of 12 hours.

We believe these findings from Teney et al. (2017) would help our future work in both Natural and Retinal domains.

7. Conclusions

VQA has a pivotal role in computer vision and natural language processing that prerequisites a system to perform much more than a single task. An algorithm that can give a right answer to an arbitrary question is the goal of artificial intelligence in our daily life.

In this thesis, we examine the danger of DR: the longer a man has diabetes, the higher his or her odds of treating DR as it is the primary source of visual impairment in individuals matured 20 to 74. We, hence, are motivated to build a computer-aided disease diagnosis in retinal image investigation with thu purpose of easing mass screening of population with diabetes mellitus and help clinicians in using their opportunity more productively.

This thesis introduces a multimodal fusion scheme WMLB between image and question data streams in both Natural and and the very first Retinal VQA. WMLB is evaluated on the most recent VQA dataset reaching state-of-the-art among bilinear approaches.

8. Acknowledgments

First and foremost, I would like to extend my deepest appreciation to my advisor, Raphael Sznitman, for his great guidance and vision. Also, I am thankful to Remi Cadene, one of the authors of MUTAN, for his enthusiastic replies on his open-source VQA codes. Last but not least, I would like to thank Thomas Kurmann and Pablo Marquez Neila for their supports on Docker and Mesos.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L., 2017. Bottom-up and top-down attention for image captioning and vqa. arXiv preprint arXiv:1707.07998 .

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D., 2015. Vqa: Visual question answering, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2425–2433.

Aone, C., Okurowski, M.E., Gorlinsky, J., Larsen, B., 1997. A scalable summarization system using robust nlp. Intelligent Scalable Text Summarization .

Ben-younes, H., Cadene, R., Cord, M., Thome, N., 2017. Mutan: Multimodal tucker fusion for visual question answering, in: The IEEE International Conference on Computer Vision (ICCV), p. 3.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 .

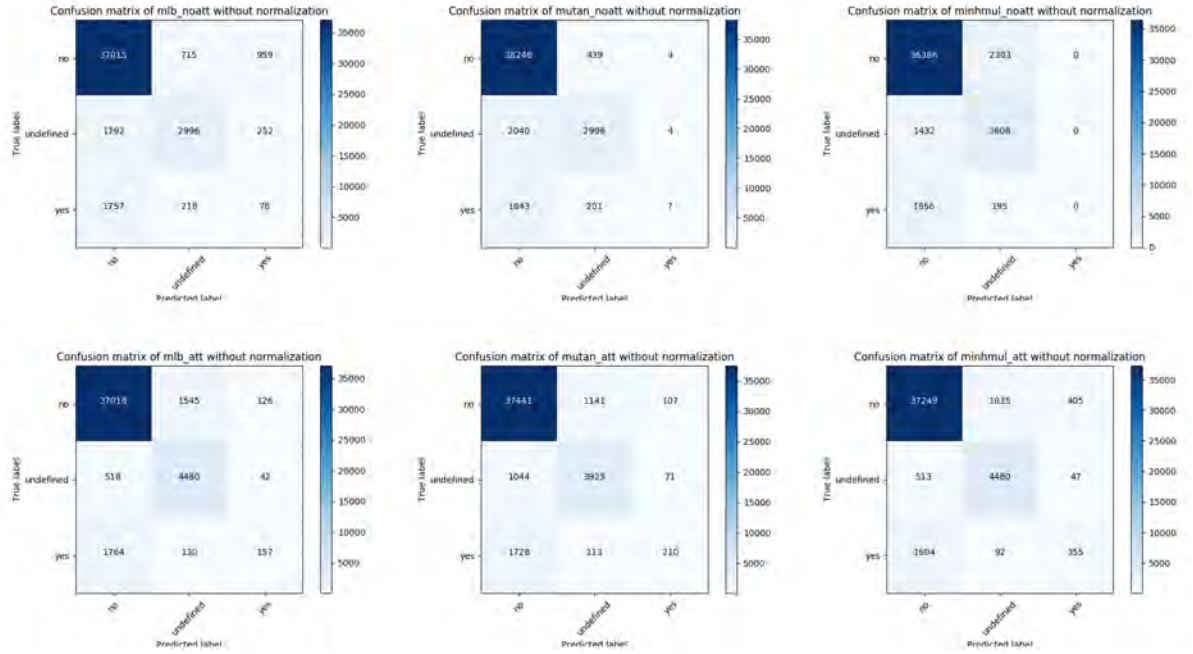


Figure 22: Confusion matrices of six methods. **Top**: without attention. **Bottom**: with attention. From **left** to **right**: MLB, MUTAN and WMLB.

Engelhard, N., Endres, F., Hess, J., Sturm, J., Burgard, W., 2011. Real-time 3d visual slam with a hand-held rgb-d camera, in: Proc. of the RGB-D Workshop on 3D Perception in Robotics at the European Robotics Forum, Vasteras, Sweden.

Finkel, J.R., Grenager, T., Manning, C., 2005. Incorporating non-local information into information extraction systems by gibbs sampling, in: Proceedings of the 43rd annual meeting on association for computational linguistics, Association for Computational Linguistics. pp. 363–370.

Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M., 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847 .

Fukunaga, K., 2013. Introduction to statistical pattern recognition. Academic press.

Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., Xu, W., 2015. Are you talking to a machine? dataset and methods for multilingual image question, in: Advances in neural information processing systems, pp. 2296–2304.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D., 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, in: CVPR, p. 9.

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.

Heymann, P., Koutrika, G., Garcia-Molina, H., 2007. Fighting spam on social web sites: A survey of approaches and future challenges. IEEE Internet Computing 11.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.

Kafle, K., Kanan, C., 2017. Visual question answering: Datasets, algorithms, and future challenges. Computer Vision and Image Understanding 163, 3–20.

Kim, J.H., On, K.W., Lim, W., Kim, J., Ha, J.W., Zhang, B.T., 2016. Hadamard product for low-rank bilinear pooling. arXiv preprint arXiv:1610.04325 .

Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S., 2015. Skip-thought vectors, in: Advances in neural information processing systems, pp. 3294–3302.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al., 2007. Moses: Open source toolkit for statistical machine translation, in: Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions, Association for Computational Linguistics. pp. 177–180.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

Lee, R., Wong, T.Y., Sabanayagam, C., 2015. Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. Eye and vision 2, 17.

Lowe, D.G., 1999. Object recognition from local scale-invariant features, in: Computer vision, 1999. The proceedings of the seventh IEEE international conference on, Ieee. pp. 1150–1157.

Lu, P., Li, H., Zhang, W., Wang, J., Wang, X., 2017. Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering. arXiv preprint arXiv:1711.06794 .

Malinowski, M., Rohrbach, M., Fritz, M., 2015. Ask your neurons: A neural-based approach to answering questions about images, in: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), IEEE Computer Society. pp. 1–9.

Meyer-Baese, A., Schmid, V.J., 2014. Pattern Recognition and Signal Analysis in Medical Imaging. Elsevier.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 .

Pirsiavash, H., Ramanan, D., Fowlkes, C.C., 2009. Bilinear classifiers for visual recognition, in: Advances in neural information processing systems, pp. 1482–1490.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Solomon, S.D., Chew, E., Duh, E.J., Sobrin, L., Sun, J.K., VanderBeek, B.L., Wykoff, C.C., Gardner, T.W., 2017. Diabetic retinopathy: a position statement by the american diabetes association. Diabetes Care 40, 412–418.

Sulong, G., Ebrahim, A.Y., Jehanzeb, M., 2014. Offline handwritten signature identification using adaptive window positioning techniques. arXiv preprint arXiv:1407.2700 .

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., et al., 2015. Going deeper

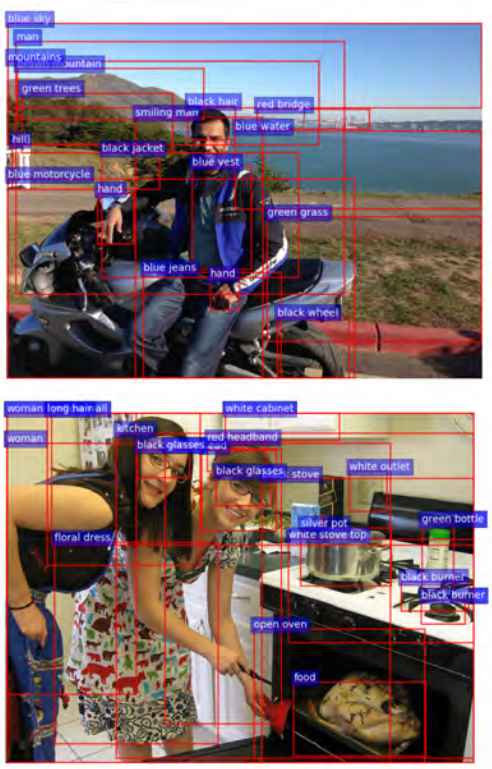


Figure 23: Example output from Faster R-CNN bottom-up attention model from Anderson et al. (2017). This figure was taken from this paper.

with convolutions, Cvpr.

Teney, D., Anderson, P., He, X., Hengel, A.v.d., 2017. Tips and tricks for visual question answering: Learnings from the 2017 challenge. arXiv preprint arXiv:1708.02711 .

Trier, Ø.D., Jain, A.K., Taxt, T., 1996. Feature extraction methods for character recognition-a survey. Pattern recognition 29, 641–662.

Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., van den Hengel, A., 2017. Visual question answering: A survey of methods and datasets. Computer Vision and Image Understanding 163, 21–40.

Xu, L., Krzyzak, A., Suen, C.Y., 1992. Methods of combining multiple classifiers and their applications to handwriting recognition. Systems, man and cybernetics, IEEE transactions on 22, 418–435.

Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., Fergus, R., 2015. Simple baseline for visual question answering. arXiv preprint arXiv:1512.02167 .