INTRA-RETINAL LAYER SEGMENTATION OF OPTICAL COHERENT TOMOGRAPHY IMAGES USING DEEP LEARNING

by

MANUEL BRADICIC
URN: 6574309

A dissertation submitted in partial fulfilment of the
requirements for the award of

BACHELOR OF SCIENCE IN COMPUTER SCIENCE

May 2023

Department of Computer Science
University of Surrey
Guildford GU2 7XH

Supervised by: Dr Roman Bauer

I declare that this dissertation is my own work and that the work of others is acknowledged and indicated by explicit references.

Manuel Bradicic
May 2023

Abstract

Since the rise of convolutional neural network architectures in the area of artificial intelligence, computer-aided systems experienced an acceleration in both research and industry. Such systems have become a hot topic in medicine, mainly used for disease diagnosis. Some researchers are even investigating such systems used for the early detection of diseases - on an almost molecular level. Moreover, a great deal of medicine has not been fully researched yet, especially using computer-aided systems, which allows researchers to explore some very niche fields of medicine. For instance, the field of ophthalmology supplies neuroscientists with images of neural tissue through optical coherence tomography imaging. In other words, this imaging technique enables to capture images of the human retina, which is essentially an extension of the brain, making it highly suitable for obtaining medical information on neurological conditions. It is a non-invasive technique, with which neuroscientists are able to obtain a large amount of data without expensive surgeries. Such data provides both neuroscientists and researchers developing computer-aided systems with a new set of tools. This Bachelor's thesis focuses on researching, developing, and implementing a model used for segmenting neurological structure in the images of optical coherence tomography. It proposes different approaches used in artificial intelligence for medical imaging segmentation and traditional methods used in computer vision.

This bachelor's thesis is centered around a provocative hypothesis "Alzheimer's disease may be an eye disease as well as a brain disease". The primary objective of this thesis is to make a valuable contribution to the field and aid the research in proving this statement.

Public version can be found on GitHub RaRetina

Acknowledgements

I would like to express my deep gratitude to my supervisor, Dr Roman Bauer, for his guidance and assistance throughout the year, supporting me with frequent meetings to attain the results accomplished in this project. Secondly, I would like to thank Dr Lilian Tang, for her invaluable support through this project, and also for giving me insights into recent machine learning technologies through my education at the University of Surrey. Additionally, I am deeply thankful for the constant support I've received from Dr Mariam Cirovic over the course of my university. Their combined efforts have significantly contributed to my academic growth and the successful completion of this project.

Finally, my sincere appreciation goes to my family and friends, who have been my source of encouragement and support when I needed it the most.

Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
AMD	Age-related Macular Degeneration
ASPP	Atrous Spatial Pyramid Pooling
CAD	Computer-Aided Detection
CNN	Convolutional Neural Network
CV	Computer Vision
DL	Deep Learning
DR	Diabetic Retinopathy
DSC	Dice Loss
GPU	Graphical Processing Units
GS	Graph Search
HMM	Hidden Markov Model
ML	Machine Learning
MAE	Mean Unsigned Error
MHA	Multi-head attention
MLP	Multilayer Perceptron
NLDA	Non-linear Deonising Algorithm
mIoU	mean Intersection over Union
ND	Neurodegenerative Disorder
OCT	Optical Coherence Tomography
RNN	Recurrent Neural Network
RPE	Retinal Pigment Epithelium
RNFL	Retinal Nerve Fibre layer
SGD	Stochastic Gradient Descent

ILM Inner Limiting Membrane

Chapter 1

Introduction

1.1 Problem background

Neurodegenerative disorders occur when nerve cells in the brain or peripheral nervous system lose function over time and ultimately die. Neurons in the brain or spinal cord damaged by these diseases cannot be repaired or replaced by the body, usually resulting in incurable conditions for the patient. Although, particular treatments may help ease off some of the physical and cognitive symptoms associated with neurodegenerative disorders, slowing their progression is not currently possible. Patients affected by these NDs can take a hit on a person's ability to move, speak or mental functions - resulting in memory loss.

Alzheimer's disease and Parkinson's disease, the two most common neurodegenerative disorders (NDs), affect millions of people worldwide. According to the Alzheimer's Disease Association in 2022, as many as 6.2 million people may have Alzheimer's disease in the United States, as well as 900,000 people in the United Kingdom, which means that 1,86% and 1.14%, of the population of those countries, respectively, are affected by this particular ND [3].

The likelihood of developing a neurodegenerative disease progressively increases with age. However, there is a growing body of research showing that dementia can be prevented, for instance, all studies put emphasis on exercise as one of the effective ways of both preventing and delaying the progression of dementia [4]. Therefore, there is a general motivation among scientists to improve our understanding of what factors play a crucial role in causing neurodegenerative disease, as well as to identify biomarkers critical for detecting early diagnosis. Being able to detect the risk of developing such a condition early in life, may result in less deliberate

and expensive ways of treating the disorder and allow early treatment and prevention.

1.1.1 The Human Retina

The retina is a photo-sensitive tissue within the eye that captures the incoming photons and transmits the information to the brain through the optic nerve. In other words, it is an extension of the brain, formed embryonically from neural tissue. This tissue contains different cells and photo-receptors which form different layers. Anatomy and histology of the human eye books [5] organise the multilayered structure into 9 distinct layers. Beginning from the vitreous body to choroid, the nine most outstanding retinal layers include the retinal nerve fibre layer and ganglion cell layer (RNFL+GCL), inner nuclear layer (INL), outer plexiform layer (OPL), outer nuclear layer (ONL), external limiting membrane and inner segment (ELM+IS), ellipsoid zone (EZ), outer segment (OS), and retinal pigmented epithelium and interdigitation zone (RPE+IZ) (See Fig. 1.1)

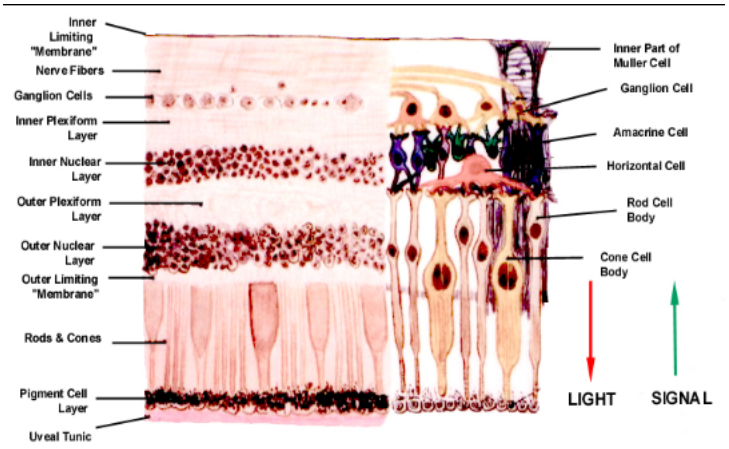

Figure 1.1: The human retina and the structure of retinal layers

An ordinary way of capturing the retinal fundus - the inner-retinal structure, is by using ophthalmoscopy. Such examination supplies ophthalmologists with a rich visualization of the macula, fovea, optic disc and retinal vessels, which are used in diagnosing various vision impairments. However, in order for clinicians to establish a more accurate diagnosis and to track the progress of impairment, clinicians need to visualise the morphological changes in the structure of the retina. In order to do so, one needs to use Optical Coherence Tomography imaging.

1.1.2 Optical Coherence Tomography

Since it was first developed in 1991 [6], Optical Coherence Tomography (OCT) is one of the most popular and powerful optical imaging techniques. Unlike conventional histopathology which requires removal of a tissue specimen and processing for microscopic examination, OCT is capable of visualizing the internal structure of biological tissues at near cellular resolution (See Figure 1.2). OCT is analogous to ultrasound imaging, except that it uses light instead of sound. It provides a tomographic (cross-sectional) imaging of tissue structure on the micro-scale *in situ*, non-invasively.

OCT technique uses coherent light to capture micro-resolution tomographic (cross-sectional) images within an optical scattering media - biological tissue. With the high resolution that it offers, OCT is used to visualise and diagnose multiple retinal diseases, such as age-related macular degeneration (AMD) [7], diabetic retinopathy (DR) [8], glaucoma [9], as well as neurological diseases, such as Alzheimer's disease [10]. Finally, changes in OCT measurements have been used to study the progression of specific neurodegenerative diseases, suggesting that the data about the neurological structure obtained by this imaging technique may be a useful biomarker for diagnosing NDs [11].

With recent development and robustness of lower cost and 'pocket' portable OCT instruments, there is a great potential and need for globally available Computer-Aided Detection (CAD) systems for instant diagnosis [12]. The use of CAD systems has been in place for several decades and has significantly contributed to the field of ophthalmology and increased the chances of correct detection, assisting specialists-readers in diagnosing and monitoring various diseases.

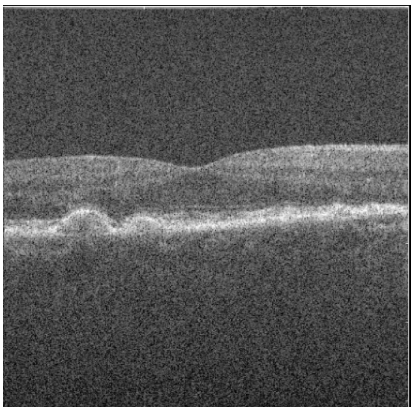

Figure 1.2: OCT image of a patient suffering from an eye disease

1.1.3 Computer-Aided Diagnosis Systems

Retinal images have proven to be crucial not only for diagnosing ocular disease but also identifying neurological disorders, as evidenced by Varghese et al. [13]. Likewise, changes in OCT measurements have been used to study the progression of specific neurodegenerative diseases, suggesting that the data obtained by this imaging technique may be a useful biomarker for diagnosing NDs. Segmentation of the neurological structure is one of the usual practices used to obtain such information. The same can be achieved in two ways: manual segmentation and automated methods. Manual segmentation is not only immensely demanding for clinicians, but also extremely expensive and time-consuming. Moreover, structural boundaries between ophthalmologists yield subjective results [14]. On the other hand, most of the already fully-functional automatic segmentation software is provided within OCT devices used only by clinicians, which intellectual property is strictly closed-sourced. An example of such a company is SPECTRALIS OCT Heidelberg Engineering (Germany). The main limitations, such as the intellectual property of medical data, in developing such systems are discussed in Section 2.1.

1.2 Aims and Objectives

The aim of this bachelor's thesis is:

- ⊙ To research, review and implement a model that is capable of segmenting the neurological structure of the human retina in OCT images collected in the UKBioBank using a deep-learning architecture with potential use in analysing subject-specific health conditions

The objectives of this bachelor's thesis are:

- ⊙ To explore and find an open-source dataset with labeled neurological structure segmentation in OCT Imaging
- ⊙ To study traditional computer vision algorithms for retinal feature extraction and quantitative thickness measure of neurological layers
- ⊙ To investigate the effect of common optimisation techniques - fine tunning, for increasing the performance of neural networks, e.g. data augmentation, annealing learning rate scheduler

- ⊙ To implement state-of-the-art computer vision architectures for neurological structure segmentation in OCT Imaging
- ⊙ To compare the proposed method with other common methods, such as traditional CNN, currently proposed in the literature
- ⊙ To create the proposed method open-sourced available for research to extend and customize for various biomedical applications, such as Alzheimer's disease

1.3 Organisation of the thesis

The structure of this report is as follows: Section 2 provides a comprehensive introduction to the background of the problem which is studied in this thesis, outlining recent advancements of deep learning in general and with applications in OCT imaging. Next, Section 3 delves into the problem analysis and planning phase, discussing the datasets employed and the techniques utilised in this study. The actual implementation of the algorithms is detailed in Section 4. Section 5 presents a thorough evaluation of the outcomes and the results achieved through this study. This is followed by Section 6 which offers a conclusion outlining potential improvements and future work. Lastly, Section 7 discusses ethical considerations that were adhered to throughout this project.

Chapter 2

Literature Review

2.1 Machine Learning in Medical Imaging

In order to design and develop ML-based CAD systems, researchers are dealing with several limitations. To make a medical dataset available for research, a special procedure has to be undertaken (See Fig. 2.1). The most pertinent point is that researchers are typically not located within a hospital and therefore do not possess access to medical imaging data. Making such data available for research is rather challenging, especially if the data has to be open-sourced to the public. The majority of Machine Learning (ML) algorithms and Artificial Intelligence (AI) models require large datasets in order to train properly and produce high-accuracy results.

Another interesting point to consider is that an institutional review board is needed to evaluate the risks and benefits of the study. Information related to past, current or future physical or mental health that reveals information about that person's health status is defined as 'Personal data' and protected by General Data Protection Regulation (Article 4(15), GDPR). This regulation on its own is slowing down the process of collecting large datasets which are essential for ML models to perform well. After ethical approval, relevant data needs to be safely stored and de-identified[1] [16]. Moreover, for open-source research, like this project, an additional effort has to be made in removing free-form annotations that occur on medical images during screening and which cannot be automatically removed. However, it is crucial to acknowledge the potential re-identification attacks - intruders, that could be launched on such datasets. Work conducted by Emam et al. [17] explains the k-anonymity, an approach for protecting privacy, yet

[1] "De-identification" is the general term for the process of removing personal information from a record [15]

over-anonymization results in immoderate data distortion, making the information less valuable for analysis.

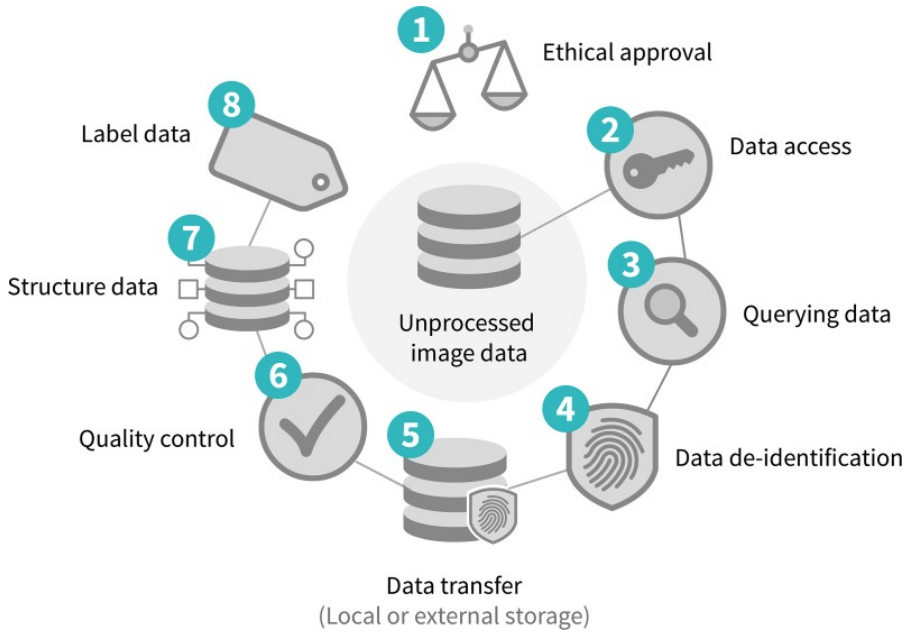

Figure 2.1: Process of medical image data handling [1]

If we analyse another problem beyond GDPR protection, we find that a retinal scan is a biometric identifier and can act as an 'identity signature', and thus is protected by Health Insurance Portability and Accountability Act (HIPAA). For instance, using head and neck Computed Tomography (CT) or Magnetic resonance imaging (MRI) imaging which is a set of sequential data, similar to OCT imaging, and with the use of volumetric acquisition, a three-dimensional reconstruction can be achieved to identify the patient. Thus, making such data publicly available requires taking potential biometric signatures into consideration [1]

Most DL models for image classification are trained on two-dimensional images with between $300 \cdot 300$ and $500x500$ dimensions, in particular, medical images have generally higher resolutions which result in requiring high computational power. For instance, convolutional neural network (CNN) architecture ResNet50, pre-trained on ImageNet uses resized images of 232x232. Training on higher image resolution is possible, however, it introduces several other issues. Firstly, in order to take advantage of transfer learning and use pre-trained architecture, which comes useful when training on a small dataset, one should preferably use the same image dimensions. Secondly, downsampling[2] of an image is an alternative, however, the reduction in spatial space

[2]The process of reduction in spatial resolution while keeping the same two-dimensional (2D) representation

could remove some potential important biomarkers in an image; examples include digital mammography as well as OCT segmentation - which by its nature is very noisy information even in the full resolution (See Section 3).

Current state-of-the-art (SOTA) algorithms for image classifications and segmentations are mostly based on a supervised learning approach, which implies that ground truth annotations have to be defined and linked to the image. The process of labelling images, i.e. defining ground truth, depends on type of problem that is tried to be solved and includes structured labels, image annotations, and image segmentation.

2.2 The Rise of Deep Learning

The idea of artificial intelligence was proposed early in the 1950s and 60s, however, it was only in 2012 that the field of artificial intelligence, then more than half a century old, had its breakthrough moment with the proposal of ImageNet - a large-scale hierarchical image database [18]. Following this milestone, the same year a team from the University of Toronto proposed the first deep neural network called AlexNet, an eight-layer CNN which emerged as the winner of the ImageNet 2012 competition becoming a game-changer for AI and CV. First CNNs were proposed as far back as the 1990s but experienced their boom only in the early 2010s. The research proposed two primary factors. Firstly, it was only in 2012 that research was furnished with unprecedentedly high-quality training data: the ImageNet database. The ImageNet database was one-of-a-kind and consisted of over fourteen million images with more than twenty thousand diverse categories. Deep CNNs had the ability to efficiently memorise and train on a such big dataset. Secondly, the achievement in hardware technology, particularly graphical processing units (GPUs), designed computational problems trained in parallel, proved to be perfectly suited for training CNNs [19].

Deep neural networks (DNNs), especially CNNs, have been dominant in the area of computer vision in the past decade, due to their ability to learn and create feature vectors of images through their convolutional layers. This process gives rise to the sense of images, building meaningful representations used for computer vision tasks. Figure 2.2 represents the progression of SOTA architectures and their performance on ImageNet. Although the recent advancements in deep learning have been astounding, there still exist various challenges to CNNs' application in medical imaging [20]. In essence, deep learning is often associated with a 'black box', since researchers

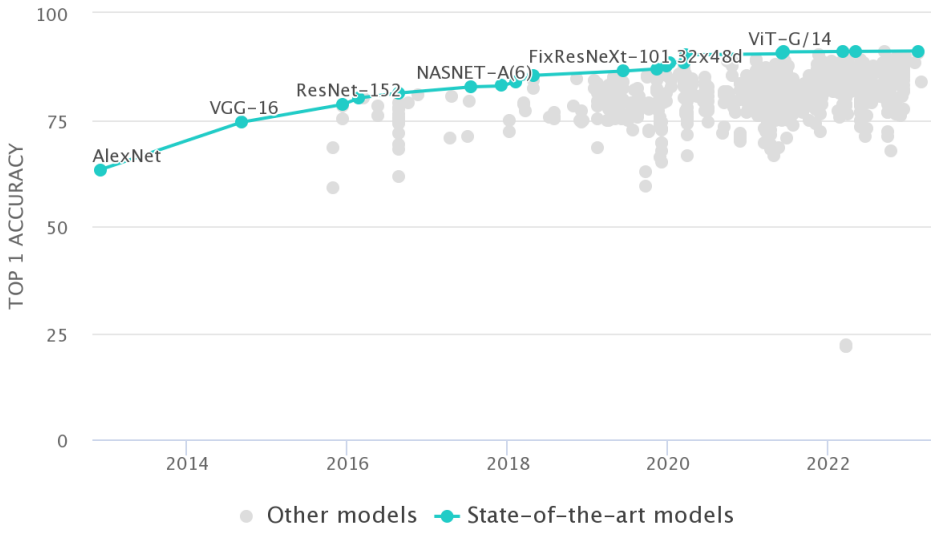

Figure 2.2: Top 1 Accuracy of computer vision architectures on ImageNet

are still trying to clarify its decisions, known as explainable AI. Another point worth considering is that only several methods are able to perform on smaller datasets, however, in the case of deep learning well-annotated large medical datasets are still an essential part of the training process. Yet, that is a very demanding and expensive process(Explained in detail in Section 2.1).

While the CNNs were designed for computer vision where image classification is conditioned on the spatial proximity - the closeness and relation between pixels, language has a sequential structure, i.e. a series of words. Recurrent neural networks were designed for dealing with sequential data and making sense of context, which was used for natural language processing (NLP). However, another advancement in the field of artificial intelligence was reached by a new architecture, known as Transformers [21].

Recurrent Neural Networks (RNNs) and Long-short Term Memory (LSTM) have been used to deal with the problem of sequential data in the area of natural language processing. Those architectures were performing acceptably, however, the problem that occurs to RNNs generally happens with LSTMs too, i.e. when sentences become too long they do not perform well. The problem was only solved by the introduction of attention by Google researchers; 'Attention is all you need' [21].

2.2.1 U-Net

On various medical image segmentation, U-Net architecture [22], named by the shape of its architecture, illustrated in Figure 2.3, has become the default in practice and achieved tremendous success. The original paper divided the architecture into two parts; the contracting path (left side) and the expansive path (right side). Such blocks are often referred as downsampling and upsampling by the literature, moreover, the literature also often suggests the middle part as the bottleneck. The main characteristic of the downsampling part of the architecture is its 3×3 convolutional layer followed by a 2×2 max pooling. The number of kernels, also called feature maps, after each block doubles which essentially allows the architecture to learn the complex structures efficiently. Similarly, in the expansion section (visually right side of the architecture) each block passes the input to two 3×3 CNN layers followed by a 2×2 upsampling layer. At upsampling, each expansive block gets appended by feature maps of the corresponding contraction layer, also called 'skip connection' or 'residual connections', which could be first seen in ResNet architecture. These connections are introduced to maintain some spatial information from the image in the downsampling part.

The typical use of CNN is a classification task, where a model has to predict an output based on the input image. However, in some visual tasks, particularly in medical imaging processing, the desired output should include localization, in other words, a class label is supposed to be assigned to each pixel.

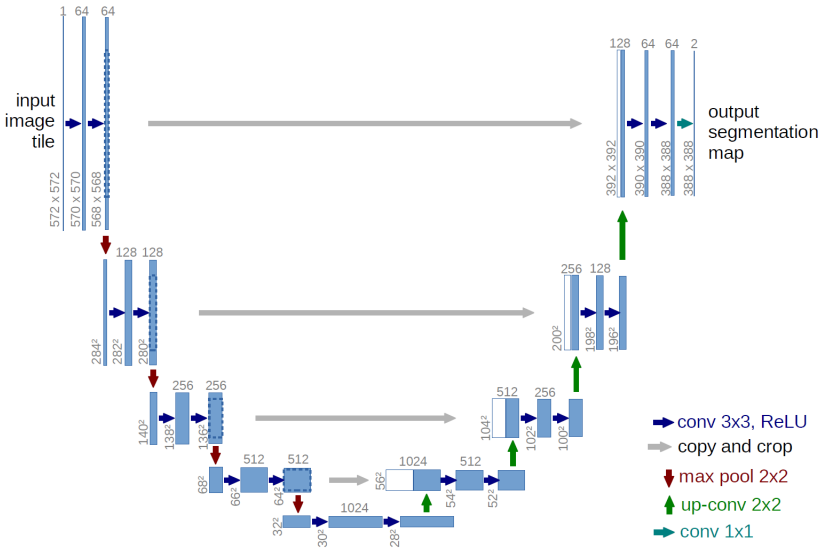

Figure 2.3: U-Net: CNN for biomedical image segmentation

2.2.2 ResNet

ResNet is a deep network proposed by He et al. [23] in the paper *Deep Residual Learning for Image Recognition*, which is one of the most influential papers in the field of computer vision. The architecture allows for learning low, mid and high-level features with depth between 16 and 30 layers. The first contribution of this paper demonstrates the importance of balancing a number of deep layers. In other words, it demonstrates that increasing the number of deep layers - stacking convolutional layers, does not necessarily increase the performance of the mode. Secondly, its importance lies in the skip connections (also known as residual connections), which were used in this paper, and could be seen presented later on in the UNet architecture. In essence, these skip connections allow the model to fit the input from previous layers to the next layers without any modifications (See Figure 2.4). Finally, another characteristic of this paper is upsampling of the previous input dimensions through identity skip connections, done by either padding zero entries for increasing dimensions or projection matched using *1x1* convolutions.

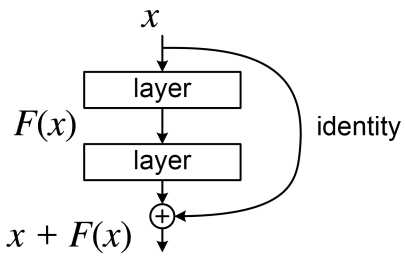

Figure 2.4: Building block of residual connections in ResNet

2.2.3 Visual Transformers (ViT)

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale is another influential paper which enriched computer vision techniques. Dosovitskiy et al. [24] showed that self-attention-mechanics, which became popular in the domain of NLP and first proposed by *Attention is all you need*, could be implemented in the field of computer vision with almost no changes to the origin transformer. This work introduces splitting images into patches, rather than looking at each pixel as a building block of an image, which would require an immense number of computational operations. Images are split into patches and together with positional encoding supply a sequence of linear embeddings which are fed into the transformer model-

encoder.

Results provided in the paper demonstrate that the variations of ViT, pre-trained on large datasets, surpassed the SOTA methodologies on some of the most popular classification datasets, such as ImageNet and CIFAR-10/100.

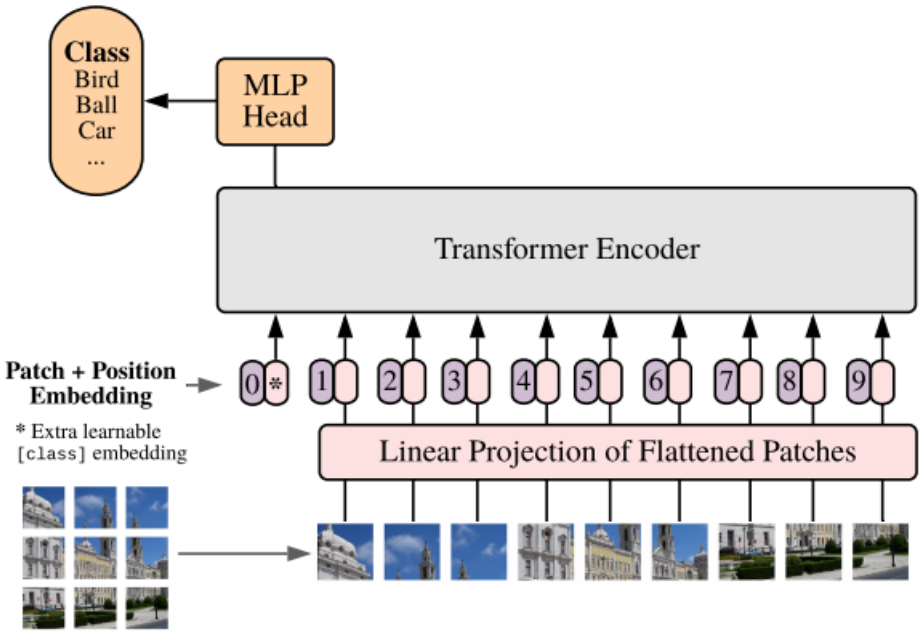

Figure 2.5: Visual transformer architecture illustrating patching process, positional embeddings and transformer encoder

2.2.4 TransUNet

Not so long after the ViTs made their way to the field of computer vision, in 2021 Jieneng Chen et al. [25] proposed the paper *TransUNet: Transformers Make strong Encoders for Medical Image Segmentation*. This architecture is composed of 3 components: transformer encoders adapted from ViT discussed in Section 2.2.3, a hybrid CNN-transformer of ResNet50 analysed in Section 2.2.2, a cascade upsampling of the hidden feature representations of the ViT, and finally UNet architecture explored in Section 2.2.1

This paper introduces CNN as the backbone of the model for feature extraction. Features extracted from the images, described as hidden features, are then fed into the 12 layers of the transformer encoder (See Figure 2.6). The need for introducing CNN is discussed in the paper by claiming that having a raw image as an input to the transformer resulted in limited localisation

abilities and restriction in giving importance to low-level details. In other words, CNN architecture is used to encode images into high-level feature representations which are patch embeddings created and fed into encoders. This allows for capturing low-level features and enhancing the fine detail of images whilst using transformer architecture. After generating a hidden feature using the transformer layer, the cascade upsampling (decoder) is used to regenerate the image from hidden into a humanly understandable representation. The same pattern of upsampling was borrowed from UNet architecture.

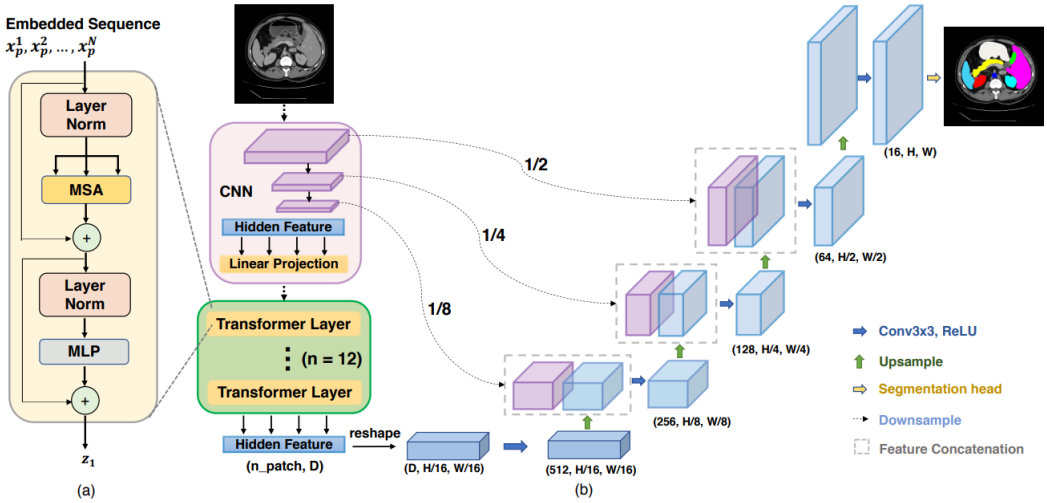

Figure 2.6: Diagram of the pipeline implemented using TransUNet architecture

2.3 Advancements of Machine Learning Techniques and Deep Learning in Optical Coherence Tomography Imaging

In neurology, detailed cross-sectional imaging using OCT has equipped both neurologists and ophthalmologists with new biomarkers, as it has been shown that thinning of particular neurological layers, in particular, retinal nerve fiber layer (RNFL) is associated with several NDs [26]. As mentioned in Section 2.1, manual segmentation is very costly and time-consuming process. For OCT segmentation, multiple studies have addressed this problem. Therefore, having a trustworthy and reliable feature, extraction methodology is of great importance for the further progress of AI in biomedical applications. In general, current SOTA OCT segmentation implementations can be categorised into two sections [27]: using artificial intelligence systems, such as CNNs and fully connected CNNs (FCNs) and mathematical-model-based methods. Mathematical model-

based methods construct a fixed or adaptive model based on previous assumptions of the retinal structure [28], sparse high-order potentials [29] and 2D/3D graph-based methods [30]. On the other hand, machine learning-based solutions, mostly deep learning, construct layer segmentation as a classification problem (binary problem), where features are extracted from the images and trained to define the boundaries.

In 2017, Fang et al. [27] presented a novel framework which composes of two main parts the first CNN layer boundary classification and the second graph search (GS) layer segmentation based on the CNN probability maps. Firstly, the CNN-GS method implements a CNN architecture to create feature vectors of specific retinal boundaries and train a classifier to delineate eight neurological layers (9 boundaries). Secondly, GS method is applied on top of the probability maps generated by the CNN to obtain the final boundaries.

Although, this study proposes a good solution, the model has been mostly trained on images with AMD structure, thus it might not perform well on healthy individuals. It achieves a 0.50 mean difference in total retinal segmentation, which suggests that it deviates from the golden standard. Finally, the authors suggest that CNN-GS is computationally intensive, therefore they are still not being widely used, especially in clinical practices.

Ghloami et al. [31] introduced a different algorithm for delineating the location and thickness of individual retinal layers. Ghloami uses a modified active contour model based on the active contour without edge method. The initialisation problem is modelled as Hidden Markov Model (HMM) and then uses the Viterbi dynamic programming algorithm to find the optimal solutions. The energy function is used then and a prior term is applied to specify active contour for OCT images. A wallet-based noise removal technique was used in this work, claiming to preserve the image quality, unlike other noise removal methods. The authors state that the average dice index similarity on the layers is 90.25 ± 0.99

Pekala et al. [32] proposed a novel segmentation method using the 103 layers DenseNet-FCN architecture described in Ref. [33], and a Gaussian post-processing regression. Furthermore, during training, the study minimized the pixel-wise cross-entropy loss using the stochastic optimiser Adam [34] with a learning rate of 1e-3. The study successfully segmented 10 intra-retinal layers with achieved mean unsigned error (MAE) of 1.06, while other methods investigated in that paper achieved MAE of 1.17-1.81 and human 1.10. The mentioned comparison was done using the publicly available University of Miami dataset [35]. This dataset was used for train-

ing, however, of the nine patients available for training, one patient was reserved as a validation set. The proposed performance displays the effectiveness of coupling DenseNet-based semantic segmentation with Gaussian post-processing regression.

In addition, Qioliang Li et al. [36] proposed an automated segmentation method based on a deep neural network named DeepRetina. This methodology uses Xception65 to extract and learn the characteristics of 10 neural layers. In the training process, this study uses an expert-labelled retinal OCT image database (which is not yet publicly available), low contrast between the layers, and the effect of speckle noise. DeepRetina consists of the following steps; Firstly, improved Xception65 is used as the backbone network to extract and learn the retinal layer characteristics, then combined with Atrous Spatial Pyramid Pooling (ASPP) for reinforcement learning and generating retinal multiscale feature information, an encode-decoder module is used to recover retinal information to capture layers and optimise segmentation results. The encoder chances the ASSP module, which uses atrous convolution with different rates to detect multi-scale convolution features (from Xception65). The model was trained on a large corpus of data - almost eight thousand images including control subjects (healthy retina) and patients with AMD and DR. Finally, automatic segmentation is achieved using the TensorFlow DL framework. This study evaluates performance using intersection over union (mIoU, 'm' stands for mean), which is a standard measure used for semantic segmentation, as well as the Sensitivity (Se) measure. According to the paper, both measures were used to compare results with existing methods. Furthermore, the reported results surpass previously mentioned studies [27, 31, 32], achieving mIoU up to 90.41% and *(Se)* up to 92.15%. Finally, all ResNet101, Inception-v3 and Xception65 were tested, however, Xception65 performed the best.

Another study, conducted by Kugelman et al. [37] trains the algorithm using the same database as this work (See Section 3.5 for more). The authors of this work train and review Cifar CNN, Complex CNN and RNN networks to classify OCT images. Moreover, this study implements a patch-based network, previously seen in several other works [27, 38]. NNs are trained using specific-sized (height $\times$ width pixels) patches of the OCT images. Each patch is assigned to a class based on the layer boundary that it is centred upon, with classes constructed for each of the three-layer boundaries of interest (ILM, RPE and CSI) or the background class (BG) for patches that are not centred upon any of the three layers; authors experimented with 32 $\times$ 32, 64 $\times$ 64, as well as 32 $\times$ 64 and 64 $\times$ 32 sized rectangular patches. They also experimented with vertical orientation. The optimisation algorithm Adam with default parameters ($\alpha = 0.001$,

$\beta_1 = 0.9$,$\beta_2 = 0.999$,$\epsilon = 1 \times 10^{-8}$) is used for training minimise cross-entropy loss with each network. No early stopping was employed and the convergence is determined based on the inspection of the validation losses, neither was the learning rate scheduler implemented, nor was transfer learning performed. Instead, each of the networks was trained from scratch with an initial randomised weight distribution. Mean Error (ME) and Mean Absolute Error (MAE) are reported as a measurement of the performance of the model. Displayed results state that UNet architecture achieved the best results out of 4 trained, however, no statistically significant performance of architecture was noticed. Performance was not compared with any other SOTA algorithms.

2.4 Literature Review Conclusion

This section investigated existing technologies and methodologies that have been identified and applied for the domain of the problem, or similar, that is this research trying to solve. The research concluded in this section will come helpful in the oncoming work when deciding on the architecture and its implementation for solving the problem of this project.

Chapter 3

Problem Analysis

3.1 Background

The accurate segmentation of neurological structures, given by OCT imaging, is crucial for the diagnosis and monitoring of various conditions. Therefore, there exists an emerging need for completing thorough data analysis on medical images of optical coherence tomography (OCT) images, however, this rich data also brings a complex and challenging issue. OCT images usually have a resolution between 20-5 μm, over *512 x 512* pixels, which in our case articulates to *6 mm x 6 mm* in real metrics. These images are very low-level, reaching structures whose sizes are in orders of micrometres, making them visible to us humans. However, there are several challenges which arise with performing data analysis on such data.

Firstly, during the imaging process invariance in image quality occurs. OCT imaging is based on interferometry [1], thus OCT images suffer from the existence of a high level of noise. Uneven illumination and signal noise are some of the factors which greatly impact the performance of screening devices, which in the end affect the accuracy of segmentation. Respectively, the variations in images and their quality are making it difficult to develop segmentation algorithms.

Secondly, neurological structures given by OCT images are complex and irregular in terms of their shape and thickness (See Figure 3.1 (a)). Moreover, images of individuals diagnosed with some of the ocular degenerations bring further deformations and complexity to those structures exacerbating the existing challenges. Furthermore, those structures demonstrate a drastic convergence at certain locations in the eye, such as the fovea (See Figure 3.1 (b)), which further

[1] a measurement method using the phenomenon of interference of waves (light)

increases the difficulty of the task and variability of the data.

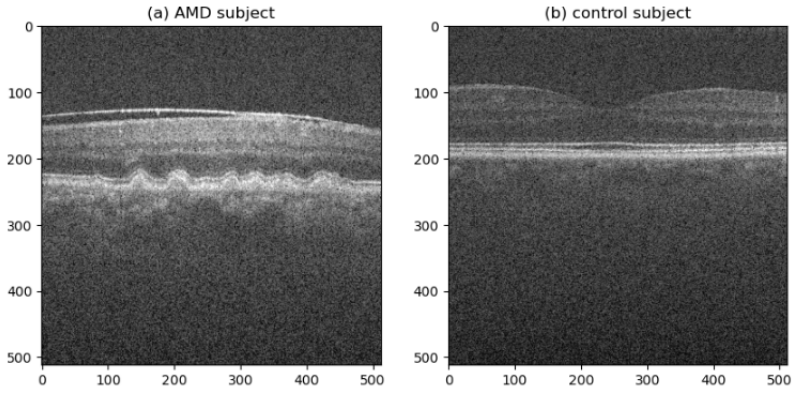

Figure 3.1: Variations in images of a diagnosed subject (a) and a healthy subject (b)

Thirdly, the lack of annotated data is one of the most challenging issues that this field of research faces. One could argue that automated segmentation predominantly relies on supervised learning, primarily due to its complexity, which entails the utilization of rich and annotated datasets. Research that works with OCT imaging is not supplied by many datasets which could be used for developing such algorithms. Studies that claim to achieve state-of-the-art (SOTA) performance rarely provide the dataset or code that was used in their research, likely due to the potential business opportunities that could arise from such work. As a result, it posed a challenge to find a publicly available dataset that includes nine layer segmentation mask. Additionally, deep learning requires a large corpus of data in order to achieve favourable results, thus out of two annodated datasets which were considered for the segmentation purpose, only one remained as an option. Further elaboration on the same topic can be found in Section 3.5.1.

Fourthly, the issue of standardization and subjectivity of manual segmentation is an additional obstacle. Manual segmentation of neurological structures requires a trained ophthalmologist which can incur significant costs. Moreover, such annotations usually provide inconsistent results across different readers, resulting in variability in datasets. Since the DukePeople dataset is a semi-automatic dataset of segmented images, this issue is less likely to be present. However, an additional concern might be the accuracy of the models utilized for segmenting such a dataset.

Finally, training deep neural network models poses difficulties such as computational complexity. As the size of the models increases, both computational complexity and the model's

performance also grow. Thus, most of the large deep learning models' parameters are in orders of millions, leading to extensive processing duration.

3.2 Data Pre-processing

Ordinarily, the analysis of OCT images is divided into the pre-processing stage, such as denoising, data augmentation, etc., and the layer-segmentation stage. Given the nature of OCT imaging technology, intra-retinal images possess an adequate amount of noise, also called speckle noise. Visual noise in images makes segmentation much more challenging since it reduces the image quality. Therefore, experimenting and implementing an effective noise-removal technique for OCT image analysis is of great importance.

To attenuate the noise contained in the OCT imaging literature suggests many techniques. For instance, Sousa et al. [39] use a bilateral filter, proposed in [40], to highlight the edges. It is a non-linear smoothing filter which preserves the edges and blurs other regions. Next, work conducted by Gholami et al. [31] mentioned the wavelet-based noise-removal, well-known among the researchers, suggesting its effectiveness in both noise attenuation as well as preserving the image quality. Some other examples of noise removal methods used in other literature include diffusion filters [41] and mean filters [42].

This work augments the dataset by applying synthetic alterations to individual samples, resulting with previously unseen samples, essentially increasing the total number of images that models work with [43]. Data augmentation is a process particularly used in computer vision (CV) tasks to increase the variety and quality of trained models. The typical motive for alterations is to result in synthetic samples, using different visual techniques, that are still representative of the same segmentation, but different to any of the existing training samples. The goal is to expose the model to diverse inputs that it sees and learns from. If done properly, data augmentation increases the variety of data beyond what the model is capable of memorizing, resulting in the model being forced to rely on generalization, rather than remembering the training set and overfitting. It is crucial to mention that the data augmentation should be performed only on the train set, which means that the validation set and the test set remained the same to preserve the original data. However, denoising could be applied to images from training and validation.

3.3 Technical Development Plan of Neurological Segmentation Algorithm

Recent advancements in deep learning, especially in the field of computer vision, have been articulated in Section 2.3. This work investigates some of the current SOTA architectures used for segmentation in the field of medicine. Vanilla UNet, previously discussed in Section 2.2.1, was defacto a rule when it came to medical imaging segmentation. Current research and findings suggest that the newer deep learning models with UNet as a backbone, outperform vanilla UNet. Therefore, a variety of models will be investigated. It is crucial to note that some of the proposed SOTA models and research for OCT segmentation may not be evaluated most studies do not share their resources with the research.

Transfer learning is a technique that allows one to use models with pre-trained weights. This technique brings great value when one is working with limited labelled data. Most of the publicly available weights are pre-trained on ImageNet dataset. The intuition behind transfer learning is that it narrows down the spacial size of the problem, given by the minimization for the loss in the multidimensional space. In other words, this approach enables a model to use pre-determined weights of the parameters, resulting in reduced training time compared to starting the training time process from scratch.

Finally, during the training process, fine-tuning will be investigated too; different optimization algorithms will be tested, batch sizes, image resolutions, etc. Models' performance is going to be evaluated using the Duke People Dataset, and qualitative and critical thinking about models' performance on UKBioBank will be given as well.

3.4 Technical development plan of the algorithm for quantitative measurement

Apart from researching, developing, and implementing the deep learning models for detecting the neurological structures in retinal images, this work will also attempt to develop an algorithm for evaluating the quantitative thickness of neurological structures provided by the segmentation algorithm, demonstrated in Figure 3.2 (last step). After generating segmented images, a quantitative algorithm will be used to detect the segmentation and produce quantitative measurements.

This can be done by calculating the distance between lines, counting pixels and transforming their values into a real metric representation. The algorithm may need to perform noise filtering and reduction in order to remove any remaining noise that might have been incorrectly predicted by the segmentation algorithm and that might negatively impact the performance of this algorithm.

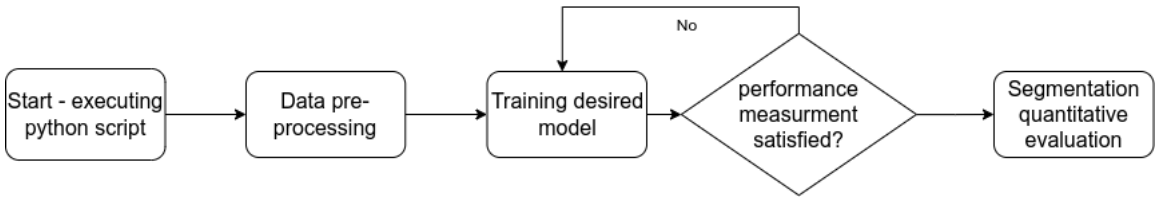

Figure 3.2: Planned workflow of the project

3.5 Datasets

As per the Problem Analysis 3.1, to meet the project's goals and aims two data sets are used, however, only one was used for training and evaluating segmentation. The first dataset, Duke People, is used to train and validate the models' performance. The second dataset, UKBioBank, is used for qualitative analysis due to its research relevance, also future work is explained in Section 6.

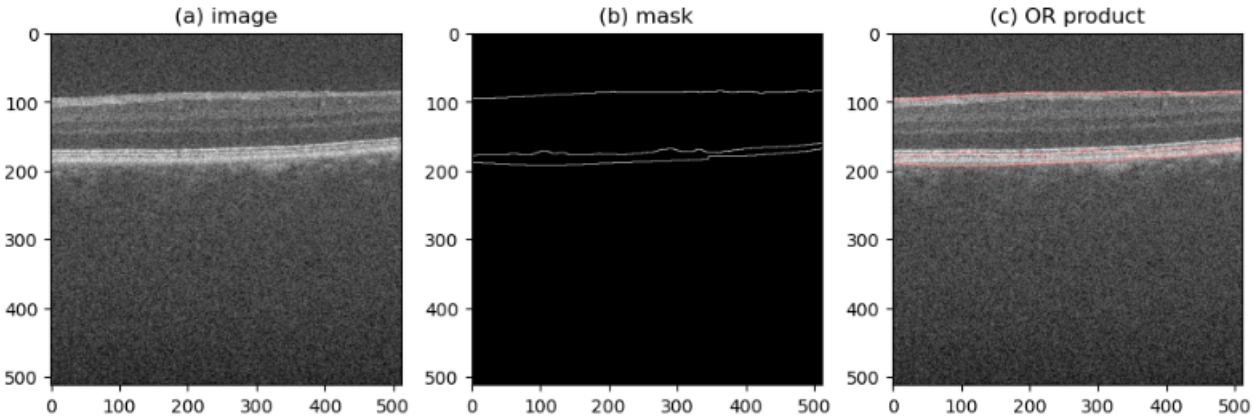

Figure 3.3: A sample from Duke People database (a) SDOCT image and (b) image mask with delineated boundaries

3.5.1 Duke People

The first dataset used for the training and validation process of computer vision techniques was presented by Duke University. Sina Farsiu et al. [44], proposed a database of one eye from 115 controlled subjects without AMD and 269 subjects with intermediate AMD. The database consists of semiautomatically delineated retinal pigment epithelium (RPE) and RPE drusen complex (RPEDC, the axial distance from apex of the drusen and RPE layer to Brunch's membrane) and total retina (TR, the axial distance between the inner limiting membrane (ILM) and Brunch's membrane) boundaries. In other words, the study shares an online atlas of a total of 38400 spatial domain OCT (SDOCT) images, their corresponding segmentations, and quantitative measures (See Figure 3.3). Initially, in this particular research, this dataset was used to study average retinal pigment epithelium drusen complex (RPEDC) and total retina (TR) thicknesses for both control subjects and subjects with diagnosed AMD. Finally, there was no additional information, such as age or sex, about the participants whose OCT images were used during the training process. These further constrain the available options and possibilities that could have been pursued using this dataset.

3.5.2 UK BioBank

UK Biobank [45] is a large-scale biomedical database and research resource that is allowing researchers to conduct scientific studies. The dataset is provided to accredited researchers and it contains a vast amount of data, including OCT images, various imaging scans, genetic data, average age, biomarkers, sex and age of participants and many others. The dataset is provided to accredited researchers and it contains medical and genetic data from 500,000 participants aged between 40 and 69 years old, where fifty-three per cent of participants in the UK Biobank were women, recruited from across the UK between 2006 and 2010. The aim of this study is to enrich our understanding of the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses.

The database contains medical imaging of the human eye, both retina images and OCT images which were acquired in the period between 2009 and 2010 using commercially available SDOCT devices: 3D OCT100, Topcon. In this study, 67,321 participants (134,642 eyes) underwent OCT imaging as a part of the ocular module [46]. It is crucial to accentuate the point that this dataset does not contain annotated segmentation, which also means that no evaluation

metrics could be calculated, but is used as a qualitative reference of the model's performance, due to its biomedical importance and research value.

3.5.3 Datasets conclusion

As stated earlier, datasets like DukePeople pose a significant challenge due to the presence of subjectivity and lack of standardisation. This particular dataset, which is semi-automatic,raises concerns about the accuracy of the ground truth mask images that delineate borders of neurological structure. This property of a dataset imposes limitations on the optimal performance that can be achieved using deep learning models, as the quality of the data they are trained on is closely linked to their performance. Thus, it is crucial to address and mitigate these limitations in order to maximise performance and achieve models with high accuracy.

Conversely, UKBioBank lacks segmented annotated data, making it unsuitable for training segmentation models. However, it can still server as a qualitative measure and a dataset for future work (See Section 6).

3.6 Feasibility

3.6.1 Feasibility of the research

The reasons explaining the challenges associated with developing an algorithm for the segmentation of neurological structures in the retina were discussed in previous sections. However, it is important to emphasize that despite the difficulties involved, it is indeed achievable with the right approach, methodology, and most importantly resources. Deep Learning has revolutionised the field of computer vision by enabling the development of highly accurate and efficient image analysis. CNNs are particularly effective in processing large scaled images and successful in extracting meaningful information and patterns from images. However, such algorithms require an exceedingly large amount of processing memory in order to be generated in some reasonable time. The majority of research uses *Pytorch* as the main framework for developing such architectures. PyTorch benefits from a large community which provides a lot of support to its users. Given these advantages, it is the chosen framework for this work. Large research clusters will have to be investigated in order to perform training of the models. The university provides two large condor clusters: *ai@surrey* cluster and *cscondor* cluster, which should allow these systems

to be trained within a reasonable timeframe.

3.6.2 Functional Requirements

To achieve a successful research outcome and to meet the aims of this research, several functional and non-functional requirements were established in Table 3.2 and Table 3.1

Req. ID	Description	Dependent Requirement	Priority Level
NFR01	Scalability: the system should be able to process large volumes of OCT images without sacrificing the accuracy of the model	FR05	HIGH
NFR02	Reliability: the system should be dependable with a minimal rate of failure in segmentation	FR02	HIGH
NFR03	User-friendly: the system should produce the results in a reasonable, user-oriented and intuitive way	/	MODERATE

Table 3.1: Non-functional requirements of this project

3.6.3 Legal Requirements

It is necessary to ensure that all the necessary regulations are being followed whilst working and developing the system using medical data. The system needs to oblige to data protection and privacy laws expressed through GDPR set by the European Union. Furthermore, considering the fact that this project requires the processing of special category data such as medical information, defined under Data Protection legislation, thus this work will require ethical and governance review by the University of Surrey.

3.6.4 Project Development plan

Development of the problem which is trying to be solved in this work is a complex and iterative process involving several stages. As displayed in Figure 3.4, the general timeline involves several iterative processes. Firstly, a thorough and in-depth research analysis and background of the problem have to be investigated. This involves researching the existing literature on

Req. ID	Description	Dependent Requirement	Priority Level
FR01	Accurate segmentation: the system should be able to accurately segment the neurological structure of interest	/	HIGH
FR02	Real-time processing: the system should be able to provide the user with a segmented structure in real time, with minimal delays. This should be achieved in order to allow for further diagnosis and treatment - if the system is used in such a manner	FR01	MODERATE
FR04	Data protection: the system should obtain all the necessary security and privacy standards that could be obtained by the corresponding regulations	/	HIGH
FR05	System integration: the system should allow for a systematic integration with other pipelines that could be used in a manner such as a diagnosis	FR04	LOW
FR06	Free and open source code (FOSC): The code should be publicly available for the research community	FR04	HIGH

Table 3.2: Functional requirements of this project

the segmentation algorithm used in OCT images, but also the segmentation algorithms used in other fields of medicine (See Fig. 3.4). Furthermore, identifying different approaches that have been used for tackling this problem, as well as understanding the weaknesses is crucial for this part. Next, data preprocessing; this research work will be conducted using already collected data, therefore data acquisition will not be included in this work. This will be obtained during the 4. and 5. stage of the timetable. Noise removals have to be considered, as well as the computational power that will be required for the training processes of this work. Furthermore, in terms of segmentation algorithm development, both existing algorithms and state-of-the-art architectures have to be tested, however, one has to consider that not all research is publicly available, thus the complete evaluation might not be achieved - Stages 7 and 8. Moreover, once the model is selected, optimization and fine-tuning are going to be obtained - research suggests several optimization algorithms that could be used. Finally, validation and evaluation of the results will be conducted in the final stage of this work (Stage 8.3).

Risk analysis could be found in Figure 3.5

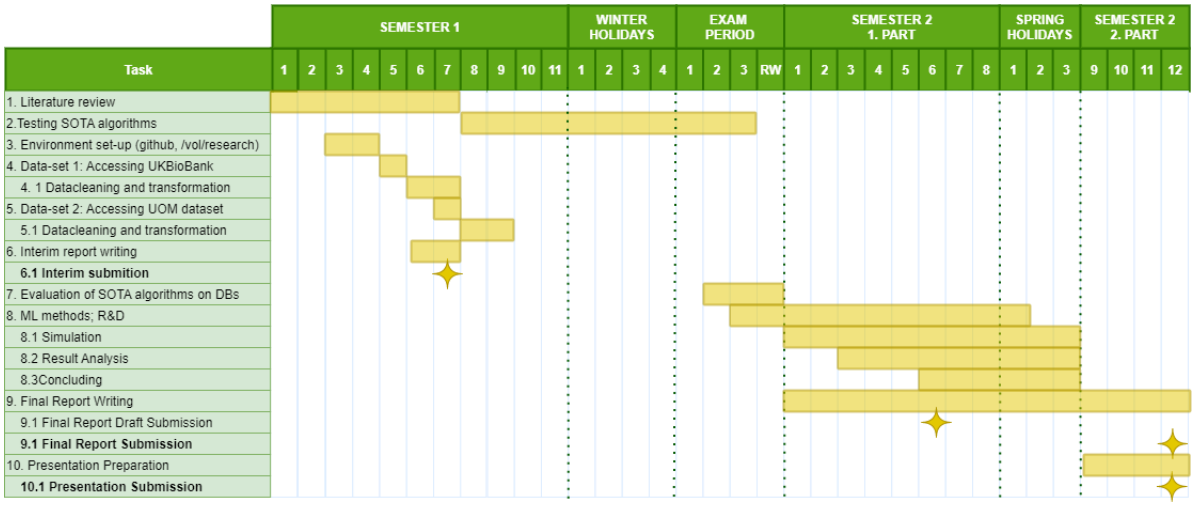

Figure 3.4: Project development timetable

Asset in Risk	Who or what is in risk?	Risk Level (likelihood, severity)	Action To Take
User Data	Data Leak; Working with different datasets and user confident data.	Likelihood: **RARE** Severity: **MAJOR**	- No images should be copied-transferred across from the university repository to private drive - Maintain the access to only anonymised data. - Strong password and 2FA should be in place for my university account - No data should be shared - Avoid connecting on unsecure Wi-Fi networks
University Machines	Computational power for image segmentation and classification might not be available or in use by other users	Likelihood: **LIKELY** Severity: **MAJOR**	- Secure AI@Surrey machines by doing the course - Understand the use of CS-condor machine - Secure an individual machine to run image classifications and computation

Figure 3.5: Risk analysis

Chapter 4

Method

4.1 Background

In this chapter, a new efficient image segmentation algorithm for extracting the thickness of different intra-retinal layers in optical coherence tomography images is proposed. The general workflow of this project is illustrated by Figure 4.1. Firstly, data loaders are created with defined transformations (augmentations), next DL models are trained using those images, and finally, generated segmentation images are further processed by the reconstruction algorithm and quantitive measurement algorithms providing a reconstructed segmentation and quantitive measures of layers.

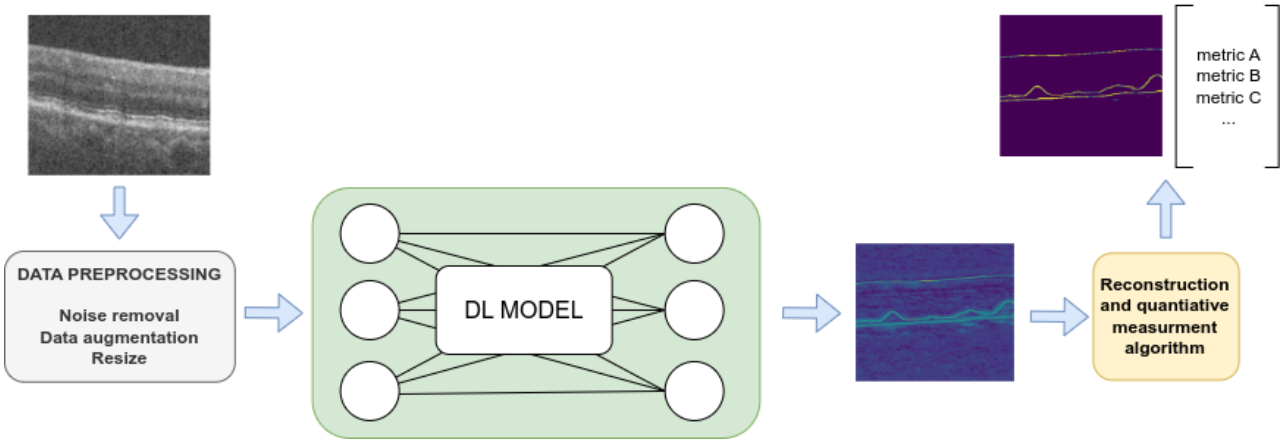

Figure 4.1: Implemented workflow of the algorithms

This work implements 3 different networks: UNet, ResNetUNet and TransUNet. Each of

these models was run using different batch sizes - depending on the size of the model. Out of the three, the UNet model is the smallest one containing just below 8 million parameters, next one ranked by size is TransUNet, a combination of a transformer, ResNet and UNet architecture which makes 67 million parameters. Finally, ResNetUNet with a total of 96 million parameters (Refer to Table 4.1 for more details).

This section explains the overall implementation of this research. It discussed the models that are being used, as well as their architectural implementation, experimental setup, optimization algorithms which were used, as well as their parameters. Moreover, it looks into cost functions which are used for calculating the accuracy of the models and their comparison with other SOTA, as well as learning rate optimisation algorithms. Finally, the implementation of the algorithm for metrical analysis is presented.

4.2 Data Pre-processing

In Section 3.2, it was emphasized that data preprocessing will be thoroughly investigated, given its potential and great influence on models' performance. To achieve the best possible results, this work dedicated significant effort to exploring those techniques.

The primary focus was on noise-cancelling algorithms. These algorithms aim to filter out noise given by the nature of the imaging device, as well as the environmental noise which impacts the quality of images. It is believed that by removing background noise, also known as speckle noise, models should be able to more easily identify important patterns and features in the data.

In addition to noise cancellation, this work also investigated various data augmentation techniques, which could enhance the dataset by introducing additional variations and increasing the size and diversity of the data.

4.2.1 Data Denoising

In general, noise is commonly regarded as a random variable with an average value of zero. Suppose one takes a pixel affected by noise, $p = p_0 + n$ of an image I, where n is the noise of that pixel. In that case, it is claimed that if we take a large number of the same pixels(N), from different images (I) and calculate their average, the expected result would be $p = p_0$, given that the noise is zero.

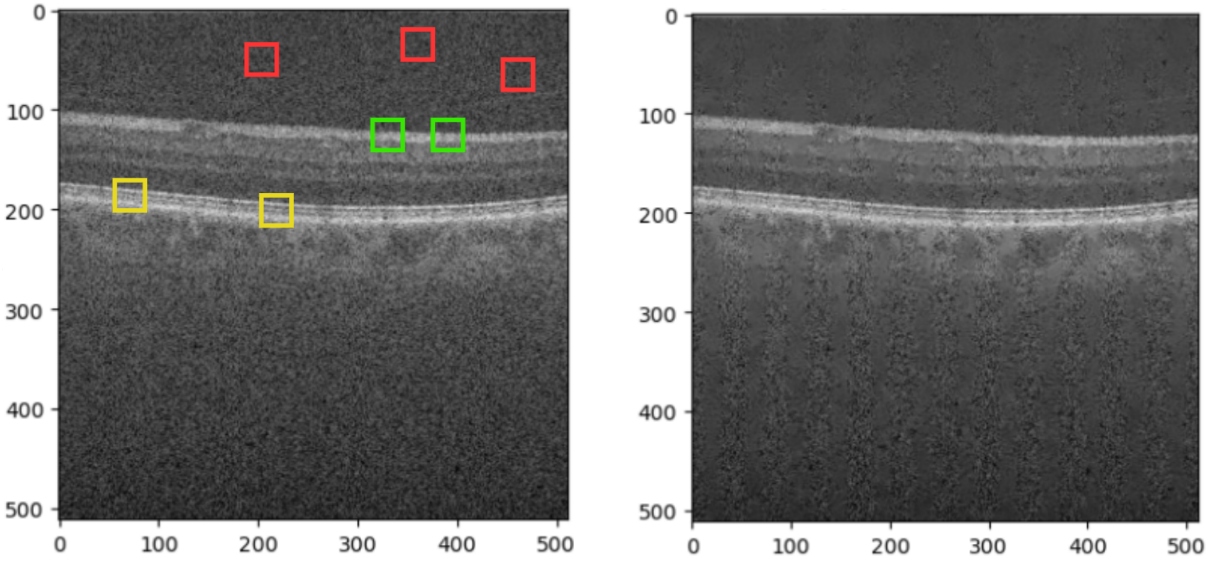

Figure 4.2: (a) principle of work of the NL means denoising algorithm (b) resulting image

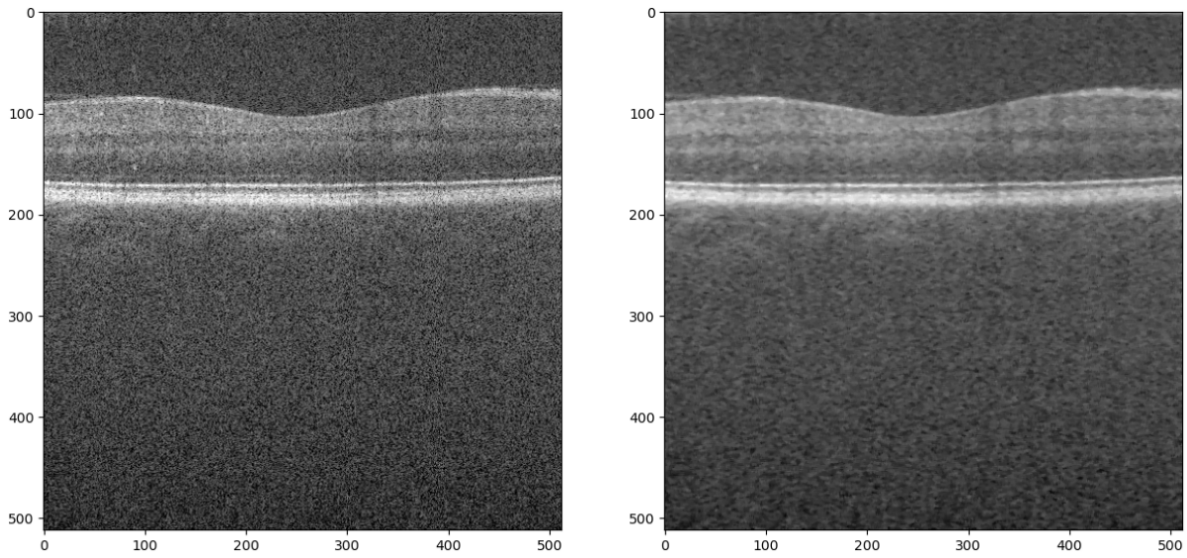

Figure 4.3: (a) raw image, (b) image after applying median filter

In Figure 4.2, the principle behind the non-local means denoising algorithm is demonstrated. The red kernels represent the noise from the region above the neurological tissue, while the green boxes represent similar kernels in the area of the Inner Limiting Membrane (ILM). Finally, the yellow patches represent similar kernels in the area between RNFL and RPE. Non-Local means denoising algorithm creates image kernels of size (m, n), identifies similar patches in the image, averages all the similar kernels and replaces the values of those kernels with the given results.

Additionally, multidimensional median filters are often used to reduce salt and pepper noise, which is a type of noise that could be used to describe OCT images. This filter works by

calculating the median value of a set of neighbouring pixels, a kernel, for each pixel in the image. The quality of OCT images can be described as delicate since some of the neurological boundaries are described by only a few pixels in the vertical axis, and although median filters can be used to remove salt-and-pepper noise effectively, they may also cause the loss of fine details that are crucial for this problem. Figure 4.3 demonstrates the performance of this algorithm.

From the given Figures 4.2 and 4.3 it is visible that speckle noise was reduced to some extent, compared to the original image. However, it is also possible that those kernels smoothed some of the important pixels with neighbouring pixels, which are crucial for delineating the boundaries. In order to determine their impact on the models' performance, both filters will be assessed and evaluated. Refer to figures A.1 and A.2 for more examples on data augmentation.

4.2.2 Data Augmentation

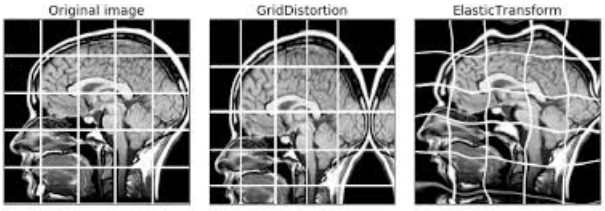

Figure 4.4: Grid distortion and elastic transform applied to an MRI image. Source: [2]

The data augmentation library used in this work is Albumentations Python Library [2]. The data augmentation used in this work consists of several parts.

Geometrical data augmentations applied to the training images and masks included: vertical and horizontal flips with a probability of 0.5; and affine transformations with rotations; Random grid shuffle with a probability of 0.3; These transformations allow the model to look at the same pictures from a slightly different perspective, in general, this could be seen as a technique which increases the overall size of the dataset which in general prevent models from overfitting.

Many authors of other medical image segmentation use colour data augmentation. However, considering the nature of OCT imaging, which provides only grey-scale images, sets some limitations. Thus, brightness and contrast adjustments with a probability of occurrence are set to 20%, to simulate variations in image quality and prevent overfitting of the model.

Finally, some recent advancements, proposed by Albumentations library [2], allow structural

transformations and deformations of images. Figure 4.4 demonstrated tissue deformation in MRI images of the brain. This approach further increases diversity and prevents overfitting of the models, it can also help the segmentation model to learn to recognise structures of interest despite tissue deformation or motion artefacts.

4.3 Neurological Segmentation Algorithm

4.3.1 Experimental Setup

Due to the computational cost of these deep learning models, all of the methods were designed to enable parallel execution across multiple GPU units. The main GPU cluster used in this work was the CScondor, provided by the university. All the experiments were run with the same computational power - 4 GPU units of RTX 6000 24GBs, in total 96GB of GPU memory was used for each experiment. To see the scripts used for scheduling jobs on clusters, check the file '*cluster/segmentation.submit_job*'.

Next, regarding the execution of experiments, each experiment was logged using the Python logger library. All the logs were printed and stored as a .txt file after every execution. The log file contained all the necessary details about model configurations, optimizers, image dimensions, used GPUs, and other relative information. Furthermore, any model which achieved the highest accuracy during the training process was saved. Finally, all the metrics used during the training process (dice loss train, dice loss validation, dice coef. train, dice coef. loss) were pickled and saved together with the models and logger file.

In terms of version control, a miniconda environment was established, enabling the public to recreate the explained experiments and proposed models.

For the segmentation algorithm the dataset which was used for training purposes - Duke people, was split into 3 sets: train, test and validation. The test set consists of 20% of the total dataset, whereas the training dataset consists of another 80% of the remaining data. In other words, the percentages for training, testing, and validation of the complete dataset are 64%, 20% and 16%, respectively. In terms of actual sizes, this translates to 9958, 3112, and 2490 samples, respectively

Furthermore, UNet and TransUNet models used the highest image resolution of *512 x 512*

pixels, apart from the ResnetUNet which used *256 x 256*. Original images have dimensions of (*512 x 1000*) however, neural networks perform better when the input dimensionality is in the order of *(m · m)*. It is worth noting that the resolution of the images has a great impact on the performance of the segmentation algorithms thus, the highest resolution is maintained and used for the processing of these models, when allowed. However, it also negatively impacts the speed of processing.

It is crucial to mention that all the images were loaded using the *DataLoader* library provided by the PyTorch community. This library allows loading images into batches, where image augmentation happens at the point of loading images, in other words, data augmentation and noise cancelling are performed at the point of loading the data, rather than preprocessed, augmented and separately stored in the database. This allows augmentation to be randomly assigned to the images, making every batch different to another - although it might be using the same images. Furthermore, it utilizes memory more efficiently by loading images on demand rather than all at once.

Model	Number of Parameters	Batch size	Time per epoch
UNet	7,782,913	8	15min
ResNetUNet	96,759,209	2	60min
TransUNet	67,865,713	32	10min

Table 4.1: Experimental properties of the implemented models

4.3.2 Neural Network Architectures

UNet Implemented UNet architecture consists of 4 downsampling and 4 upsampling convolutional steps, also known as encoder and decoder architecture. Image input of 512 gets downsampled to 256,128 and 64 convolutional layers, using max pooling of kernel size 2, followed by the upsampling process (Illustrated by Figure 2.3). For more details on layer dimensions check *architecture/unet.py* file. This model was trained from scratch, meaning that it won't rely on pre-existing weights (transfer learning).

ResNetUNet Similarly, ResNetUNet architecture maintains the characteristic of an UNet - shaped as a letter 'U'. However, on the decoder side, decoder blocks introduced by the UNet

architecture are used. This suggests that the idea of this model is to maintain the property of a ResNet, where ResNet convolutional blocks are used for creating feature maps (kernels) for identifying patterns in images, while the decoder blocks are borrowed from the UNet. Residual connections between the decoder-encoder are maintained too. The final classifier consists of a convolutional layer with a ReLU activation function, followed by another convolutional layer and a sigmoid function. Finally, this network was using pre-trained weights, trained on ImageNet, based on ResNet101 architecture provided by the PyTorch Community.

TransUnet This architecture combines several architectures, mainly, it takes the encoder part from the ResNet network, using ResNet feature maps as the input vectors for the visual transformer, which processes images in patches of (16, 16), and finally, the features maps in latent space, produced by the visual transformer, are upsampled using UNet decoder (See Figure 2.6). TransUNet architecture is not using transfer learning, which indicates that all the weights are trained from scratch.

Firstly, the ResNet convolutional layers have an input channel size of 3, and an output of 128, which means that the feature maps (usually annotated with c) have a size of 128. A kernel size of 7 was used inside the convolutions with stride and padding of 2 and 3, respectively. Secondly, the encoder of the visual transformer consists of 12 encoder blocks, where each block consists of 4 multi-head attention (MHA), a multi-layered perception (MLP) with two linear layers, of input/output dimensions of 512, and a Gaussian Error Linear Unit (GELU) activation function. Finally, the dropout rate of 10% is used. Thirdly, the decoder layer - the upsampling process, borrowed from UNet was implemented. It consists of 4 upsampling blocks and each block consists of an upsample layer and two sequential layers built of a convolutional layer, batch normalization layer and an activation function rectified linear unit (ReLU). (Further details could be found in the file *'architecture/transunet.py', 'architecture/vit.py'*)

4.3.3 Optimization algorithms

Optimization algorithms are used to train deep neural networks. They are mostly gradient-based algorithms which optimize neural network weight by minimizing loss. In other words, they are calculating the next values of weights which a network is going to use based on the gradient in a multidimensional latent space. The two most commonly used optimizers for training neural

networks are Stochastic Gradient Descent (SGD) and Adam.

SGD works by updating the network's parameters (weights) based on the gradient of the point in the latent space, described by the number of features (parameters) that each network has. On the other hand, the Adam optimizer is an alternative to SGD when training models. Adam uses a squared gradient to scale the learning rate and it also benefits from momentum by using the moving average of the gradient [47].

In this work, both optimizers are implemented. SGD parameters used are: $lr = 0.01, m = 0.9, wd = 1 \cdot 10^{-4}$, where learning rate is annotated by *(lr)*, momentum *(m)* and weight decay *(wd)*. Similarly, Adam's default parameters were used apart from the starting learning rate *(lr)* which remained the same.

4.3.4 Dice Loss - Cost Functions

A cost function is used to measure the dissimilarity between two images. In this work it was used to compare the ground truth mask and generated mask. This dissimilarity is measured using particular metrics, which are then used during the backpropagation in order to update weight and move across the latent space and find minimums. In order words, the aim of such a function is to maximize the overlap between two images. The Soreness-Dice loss, also known as Dice loss, is a common loss metric for segmentation tasks.

$$L_{dice} = 1 - \frac{2 \cdot \sum p_{true} \cdot \sum p_{pred}}{\sum p_{true}^2 + \sum p_{pred}^2} \tag{4.1}$$

The dice loss function, explained by equation 4.1, is calculated by subtracting the dice coefficient from one. In other words, is based on the ratio of the sum of correctly segmented pixels multiplied by the sum of predicted pixels multiplied by two, to the sum of the squared predicted and squared actual pixels, and all subtracted from one. Dice loss, in particular, is one of the most common functions used in medical image segmentation. Moreover, most of the SOTA algorithms proposed by the literature use dice loss.

Peformance metric While cost functions are used for weight adjustments in the neural networks, this work also integrated performance coefficients. Those coefficients essentially indicate the model's performance and accuracy, on a scale from 0 (worst) to 1 (best). The dice coefficient is one of those metrics. Equation 4.1 illustrated the computation of dice loss, which is

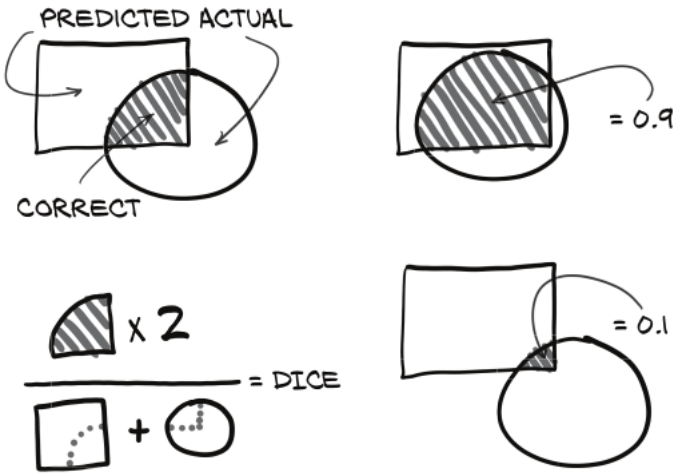

Figure 4.5: Principle of work behind Dice loss

determined by subtracting the dice coefficient from 1.

In addition to the dice metric, this work also implemented another performance metric, known as the Jaccard index, or the Intersection-over-Union (IoU). The jaccard index is similar to dice loss in measuring dissimilarities, it is defined as the intersection of the two images divided by the union of their sizes (See equation 4.2).

$$L_{jaccard} = \frac{A \cap B}{A \cup B} \tag{4.2}$$

4.3.5 Cosine OneCycle Learning Rate

One cycle learning rate is a technique of changing the learning rates of optimizers depending on the number of epochs. It was introduced by N. Smith et al. [48]. The so-called one-cycle policy anneals the learning rate from an initial low learning rate to some pre-specified maximum learning rate and then decreases to some minimum learning rate.

What this technique really does, is it makes the model weights independent of the initial data that is being loaded, allowing the learning rate to reach its maximum value just after the warm-up period, and finally, followed by a decrease. This nonlinearity of the learning rate prevents the model from overfitting and getting stuck in a local minimum too early.

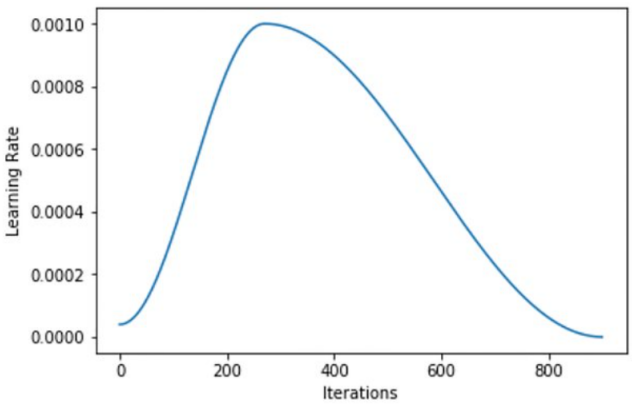

Figure 4.6: Underlying mechanism of a cosine one-cycle learning rate scheduler

4.4 Algorithm for metrical analysis

The second part of this research consists of implementing an algorithm, *'walking pixel algorithm'*, for metrical analysis (See the two final stages in Figure 4.1). This algorithm is used to generate a metrical analysis of the neurological structure, illustrated in Figure 4.7. This algorithm measures the area between the lines of a segmented image, in other words, it calculated the surface of the neurological layers, between retinal pigment epithelium (RPE) and RPE drusen complex (RPEDC, the axial distance from apex of the drusen and RPE layer to Brunch's membrane) and total retina (TR, the axial distance between the inner limiting membrane (ILM) and Brunch's membrane) boundaries.

Firstly, since the segmentation images generated by DL algorithms do not occur to be flawless, a computer vision algorithm is run to reconstruct missing parts of the borders. This was done using *OpenCV* library. Images are exposed to dilation, followed by an erosion using a horizontal kernel of dimensions $K = [1,10]$.

Secondly, the 'walking pixel algorithm' is implemented which essentially analyses the picture in a systematic way. It starts from the top left corner and walks towards the highest pixel in that column. It is looking for 3 flags, each flag is identified by the pixel value of *255*, which annotated that a border is being reached. This way, the algorithm is able to count the number of pixels for the relevant regions (denoted as region '2' and '3' in Figure 4.7). The algorithm is explained by Algorithm 1.

Algorithm 1 Pixel counter algorithm

```
1: w ← img_width
2: h ← img_hight
3: cl ← 0
4: pc ← [0, 0, 0, 0]
5: for move ∈ {0, ..., w − 1} do
6:     cl ← 0
7:     for walk ∈ {0, ..., h − 1} do
8:         p ← img[walk, move, 0]
9:         p_n ← img[walk + 1, move, 0]
10:        if p equals 255 then
11:            if p_n equals 255 then
12:                pass
13:            end if
14:        else
15:            cl ← cl + 1
16:            pc[cl] ← pc[cl] + 1
17:        end if
18:    end for
19: end for
```

Where w is the width and h is the height of an image, cl denotes the current layer which describes which out of 4 areas the pointer is in (before RPE, RPE to ILM, ILM to Brunch's membrane, and Brunch's membrane; areas 1 to 4 illustrated in Figure 4.7). Symbol p annotates the current state of the pixel, where $p \in \{0, 255\}$.

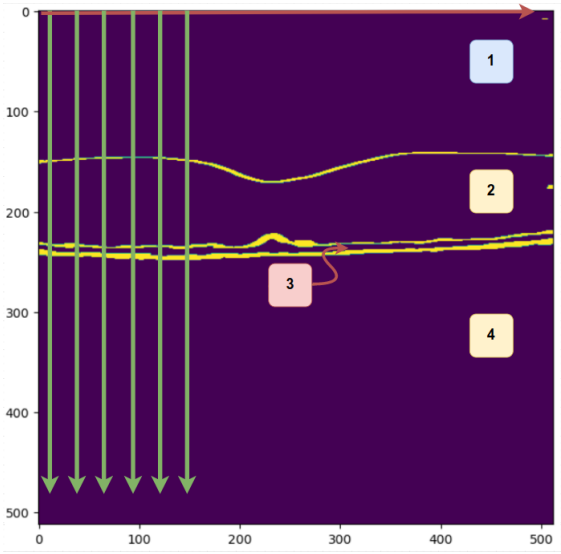

Figure 4.7: Illustration of the metric analysis algorithm

Chapter 5

Evaluation

This section showcases the results achieved in this project. It offers an in-depth assessment of the implemented models as well as a critical evaluation.

To further assess the accuracy of the model, several metrics were used to judge the model's performance: mean intersection over union (mIoU), dice metric (DSC), and GPU calculation times. Every target image in the test set already contained pixel values that are associated with a specific neurological layer, in other words, ground truth masks for the test set were provided with the dataset (Refer to Section 3.5.1).

It is crucial to note that these models are highly demanding in terms of computational resources. Despite using a total of 96GBs of GPU memory per experiment, the large dimensions of the images and the size of the dataset resulted in each model requiring dozens of hours to process. Therefore, the scope of this study was constrained by the number of computations that could be executed, given the limited resources by *ai@surrey* and *cscondor* clusters

5.1 UNet

Among all the evaluated models, the vanilla UNet model demonstrated the poorest performance across all optimization variations. This outcome was somewhat anticipated, given that all the other implemented models are an essential enhanced version of this vanilla model. The lowest dice loss achieved is 0.9625, dice metric 0.5222 and mIoU 0.3459. It is evident from Figure 5.3 that the model achieves its minimum loss at approximately the 25th epoch, and it demonstrates the signs of overfitting from there onwards. The model was running between 70-100 epochs,

where early stopping was introduced to prevent overfitting and memory consumption. On average, the model took around 10 minutes for each epoch, cumulating to a total processing time of around 13 hours. The highest achieved accuracy was achieved using Adamx optimiser, a variant of the Adam optimiser.

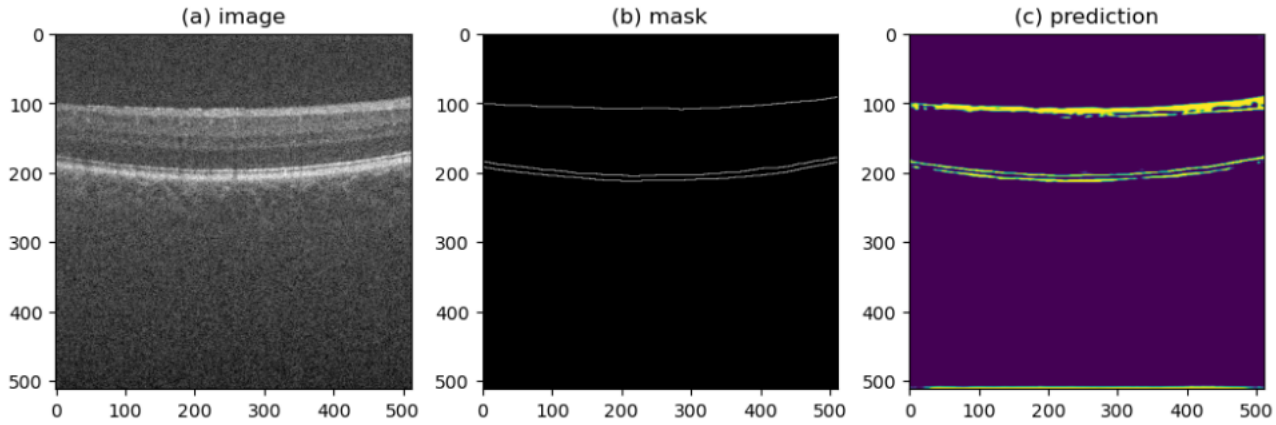

Figure 5.1: Performance evaluation of a UNet model of a healthy subject

Figure 5.1 demonstrates the execution of a UNet model on a healthy subject's OCT image. From the image, it is evident that the model is capable of extracting the features, however, it encounters some difficulties with delivering a comprehensive segmentation. There are several factors that could contribute to these results. UNet architecture, by its nature, is relatively small compared to other deep-learning models used in this work. These results suggest that UNet feature maps struggle with extracting the necessary features for recognising the layers in the model. However, it is highly probable that the challenges posed by the dataset limit the performance of the model. This work experimented with Gaussian noise removal as well as the NL means denoising algorithm, however, none of the algorithms contributed to the performance of the model. In fact, they negatively impacted the model, worsening its performance. This may be due to the fact that denoising algorithms remove some of the features crucial for delineating boundaries. Finally, the number of parameters often, but not necessarily, defines how powerful a deep learning architecture is, UNet model has just over 6 million parameters, which isn't a small architecture, however, since its proposal many newer architectures and variations were proposed.

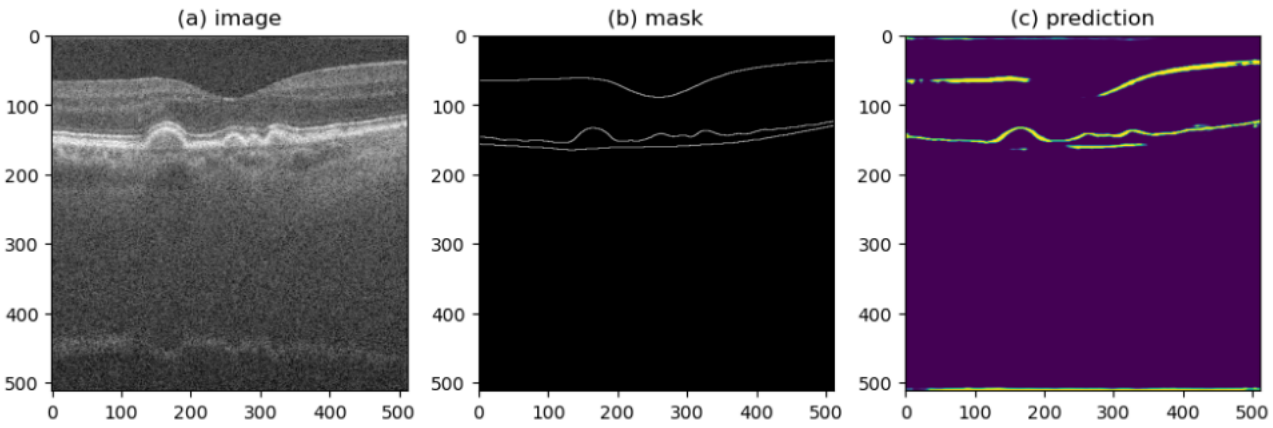

Figure 5.2: Performance evaluation of a UNet model on an individual afflicted with AMD

Figure 5.2 demonstrated UNet performance on the image of a subject suffering from Age-Related Macular Degeneration (AMD). This illness affects the central part of the retina causing progressive deterioration of the macula. It is visible from figure (a) that the neurological layers near Brunch's membrane are affected resulting in the deformation of these layers. It is visible from the figure that although the deformations occurred, the model successfully delineates one of the layers. Furthermore, this image also demonstrated the fovea of the eye. The fovea naturally occurs as a dip in the retinal surface, creating a concave dip in the OCT image, even in healthy subjects. It is visible that the model did not successfully segment the structure of the INL layer (top one). Despite being trained on several hundred images, which included the fovea of the eye, the model shows the difficulty in creating feature maps and learning the necessary patterns for accurately segmenting such a complex structure.

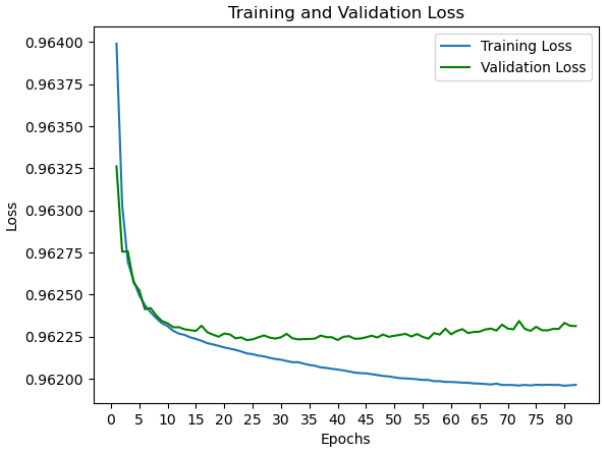

Figure 5.3: Training and validation loss for the best UNet model

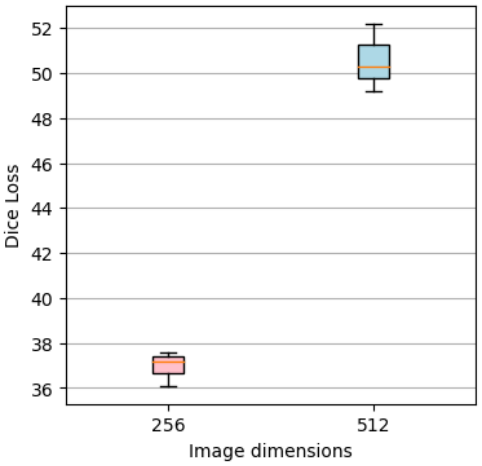

Figure 5.4: figure

Figure 5.4 showcases the performance of the UNet model with an image dimension as a variable. It is evident that the model's performance drastically increases with the image dimensionality of 512x512. This outcome was expected, as reducing image dimensionality deduces the spatial information of an image. In other words, it increases the challenge of segmenting such a complex structure. Furthermore, Figure 5.3 clearly demonstrates that the model is learning during the initial 25-30 epoch, but after that point, it enters a phase of overfitting and fails to show any improvements.

5.2 ResNetUNet

Convolutional neural networks (CNNs) like ResNetUNet often demonstrate improvements in performance when augmented data is utilized, unlike visual transformers (ViTs). By incorporating synthetic data generated through the 'albumentations' library, significant improvements have been observed. The model attained a maximum dice coefficient of 0.6581. However, it is crucial to reiterate that this evaluation was conducted on images with a resolution of 256 pixels.

Figure 5.1 demonstrates the highest results achieved finetuning the ResNetUNet model using the pre-trained Resnet101 weights as a base. The 'Train' subsection in this table demonstrates dice score results, both coefficient and loss, on the validation accuracy, during the training process. Whereas, the 'Test' subsection in the table demonstrates the same metrics including intersection over the union ('$mIoU$', m stands for mean). The baseline model displayed in the

table did not utilize any data preprocessing or augmentations and serves as a benchmark for evaluating how different preprocessing techniques impact the performance of the model. The augmented model displays the results upon the augmentation of the data during the training process, whereas median blur demonstrates the results achieved using the median blur filter and augmentation. All the models were computed 3 times and the standard deviation was calculated.

Table 5.1: ResNetUNet result table

	Train		Test		
	DSC ↑	DSC(loss) ↓	DSC ↑	mIoU ↑	DSC(loss) ↓
augmented	0.6581±0.0001	0.3939±0.0001	**0.6583**	**0.4951**±0.0001	**0.3927**
baseline	0.6569±0.0002	0.3958±0.0002	0.6572±0.002	0.4939±0.0001	0.3945±0.0001
median blur	0.6569±0.0002	0.3958±0.0002	0.6572±0.002	0.4939±0.0001	0.3945±0.0001

Upon examination of the table, it becomes evident that the augmented model outperformed the base and median blur models. Similar to the UNet model, noise reduction algorithms have had a negative effect on the performance of the model. Since this model is a CNN, it benefits from feature maps and the ability to identify patterns in images to generate meaningful outcomes. This study posits that denoising algorithms may have a negative impact on the spatial meaning of an image, reducing its spatial information, which enables CNN to create features and find patterns.

Figure 5.5 demonstrates ResNetUNet performance on 3 different images from the test set, where image (a) is a retinal image of a healthy individual and images (b) and (c) from a subject diagnosed with AMD. From the depicted illustration, it is visible that the model effectively segmented the images related to a healthy individual (a). However, the model demonstrated worse performance when it came to segmenting images associated with a condition. Finally, although figure (c) is classified with the AMD condition, the model still performed well and achieved a dice coefficient of **0.9357**.

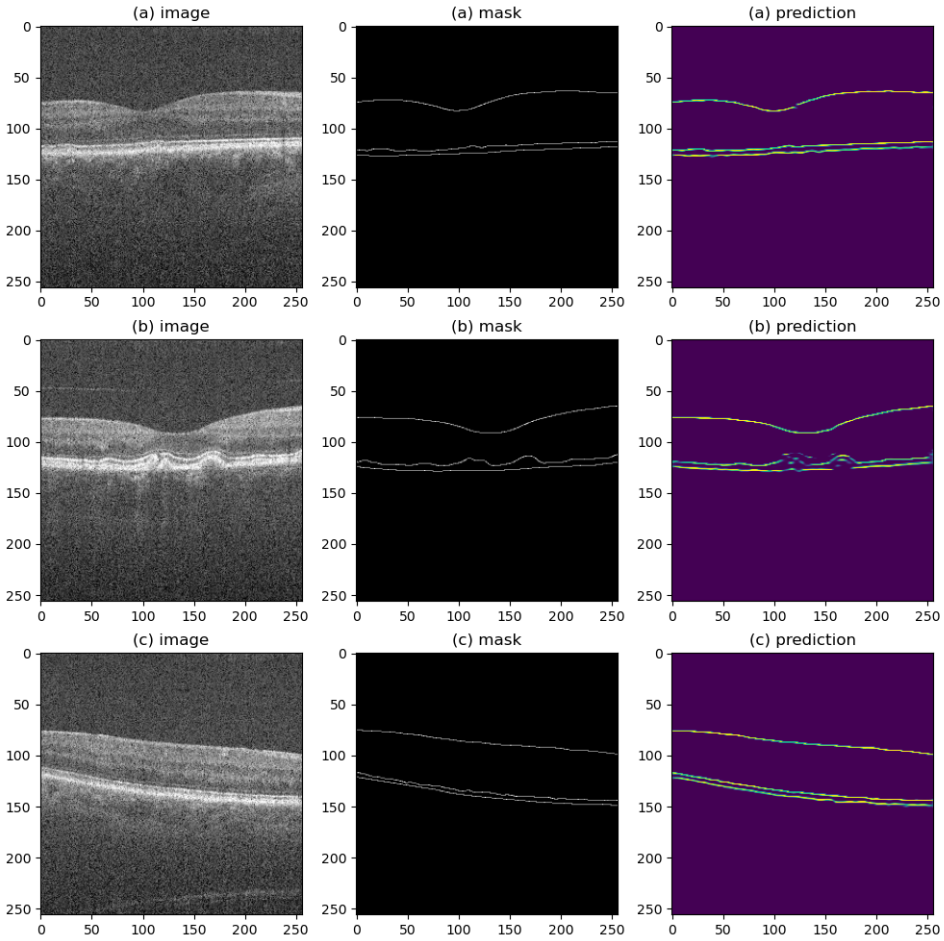

Figure 5.5: Performance evaluation of *ResNet_ augmented* model on an individual afflicted with AMD

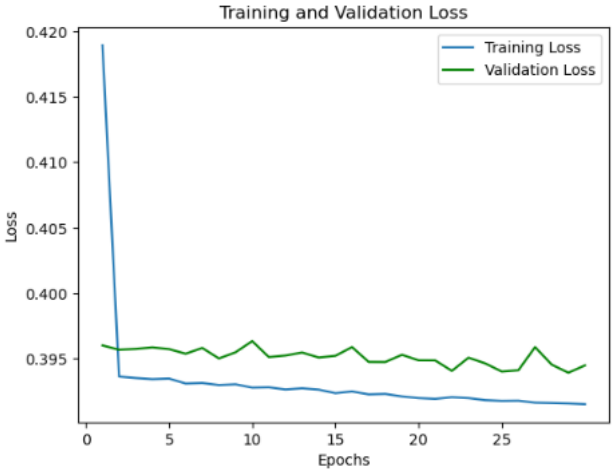

Figure 5.6: Training and validation loss of the best ResNetUNet

Figure 5.6 depicts training and validation loss during the training process of the model which

achieved the highest dice coefficient. It is visible in the figure that the model achieves relatively high accuracy on the first epoch. This is due to the transfer learning, the model's weights are already predefined for the ImageNet database, and thus a relatively small number of parameters needs to be altered. The model was trained for 28 epochs using a batch size of 4 and it was trained on 96GB of GPU memory. The total time needed to complete 28 epochs was nearly 8 hours.

5.3 TransUNet

Among all the models evaluated, the TransUNet architecture delivered the most favorable results. By its design, TransUNet architecture differs from the classical CNNs architecture due to the attention mechanism of a visual transformer implemented in this architecture and is one of the most recent SOTA architectures in medical imaging.

Table 5.2: TransUNet result table

	Train		Test		
	DSC ↑	DSC(loss) ↓	DSC ↑	mIoU ↑	DSC(loss) ↓
model_1	0.6484±0.0045	0.2745±0.0045	0.6471±0.0020	0.4803±0.0019	0.2747±0.0020
model_2	0.6837±0.0020	0.2491±0.0020	0.6839±0.0011	0.5219±0.0011	0.2488±0.0011
model_3	0.6876±0.0014	0.2432±0.0014	**0.6882**±0.0009	**0.5268**±0.0011	**0.2422**±0.0009
model_4	0.6832±0.0017	0.2477±0.0017	0.6839±0.0015	0.5219±0.0014	0.2488±0.0015

Table 5.2 presents the performance and evaluation of four distinct TransUNet models applied to the DukePeople dataset.

The first examined model, as displayed in the table, utilised Gaussian denoising filter in its preprocessing. Despite its application, the model achieved the lowest Dice Similarity Coefficient of 0.6471, among all four experiments. The model employed the Adam optimizer, as detailed in Table 5.3, thus one-cycle learning rate was not implemented in parallel.

Moving onto the second experiment, the model used the non-local means denoising algorithm, which delivered a marked improvement in performance compared to the initial model. This model's configuration includes default geometrical augmentation, SGD optimizer, and one-cycle learning rate suggesting the potential impact of these selections on the improved outcomes.

The third experiment yielded the most desirable results among the models tested, with a DSC coefficient of 0.6882 and a mean Intersection over Union (mIoU) of 0.5268. This model's configuration did not include any denoising algorithm. Instead, regarding the data pre-processing it applied geometric augmentations exclusively. Additionally, it had an increased batch size, of 24, and computed over 111epochs.

Finally, the fourth experimental configuration investigated the default SGD optimizer with geometric augmentation and yielded the same results as *model_ 2*. It is essential to note that the mean computational time for these models was approximately 34 hours, which varied depending on the pre-set batch size.

In conclusion, it is visible that the models do not benefit from noise removals, both Gaussian and NLDA. The best-achieved accuracy was derived from a combination of SGD optimiser and one-cycle learning rate, including geometrical augmentation.

Table 5.3: Implemented parameters of the 1,2,3,4 models

	Models			
Parameters	*1*	*2*	*3*	*4*
optimizer	Adam	SGD	SGD	SGD
batch size	16	16	24	24
learning rate (lr)	$1e^{-2}$	$1e^{-2}$	$1e^{-2}$	$1e^{-2}$
max (lr)	-	$1e^{-2}$	$1e^{-2}$	-
momentum	-	0.9	0.9	-
weight decay	-	$1e^{-4}$	$1e^{-4}$	-
number of workers	4	4	6	-
patience	8	8	8	6
epochs (e)	79	77	111	80
total time (h))	41	40	80	24
augmentation	$\bigoplus$	$\bigoplus$	$\bigoplus$	$\bigoplus$
noise removal	*gausian*	*NLDA*	-	-

$\bigoplus$ geometrical augmentation
$\bigotimes$ colour augmentation

Additionally, when comparing the performance with the previous two models, it is evident that the models continue to struggle with the segmentation of the images impacted by AMD. This suggests that the models achieved their maximum capacity on the given data, however, they might be constrained by the limitations provided by the dataset during the training process.

In other words, the provided semi-automatic segmented dataset, might not contain correctly annotated boundaries. Thus, limits the performance of the models to 0.68 DSC score overall.

Figure 5.7 presents the images generated by the model that yielded the most impressive results. As can be observed from image (a), the model demonstrates exceptional performance, achieving a DSC score of 0.9762 for that particular image. In contrast, image (b) achieves a significantly lower DSC score of 0.3789, illustrating the model's vulnerability in performance across images affected by AMD.

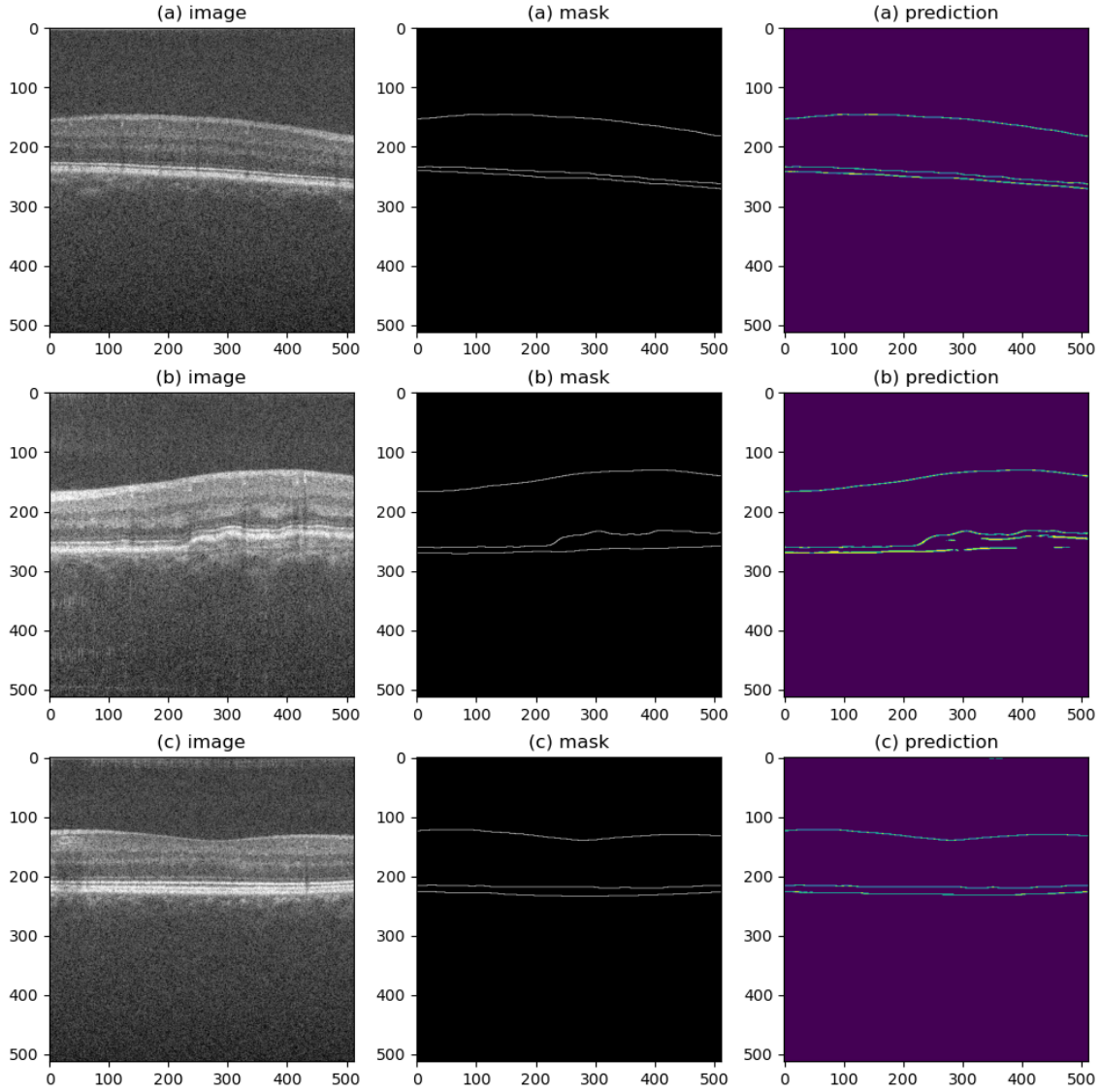

Figure 5.7: Performance evaluation of a TransUNet model on (a) control subject, (b) and (c) AMD

5.4 Comparison with Existing State-of-the-Art Methods

Most of the SOTA algorithms perform their image segmentation tasks on datasets that are not publicly available. This makes it challenging to perform a fair comparison of the results obtained from different studies, as the datasets used for training and evaluation can significantly impact the performance of the models.

However, an opportunity for comparison arises with the study conducted by [37], which presents results based on the DukePeople dataset - the same one used in this study. The results from this previous study are reported in terms of Mean Squared Error (MSE), however, despite the use of MSE for reporting results, a closer look at the images generated in that study reveals that the model proposed in this study delivers superior performance, the segmented regions in generated images of this study seem to more accurately and precisely capture the delineating area.

This suggests that the technical improvements introduced in our current model - in terms of the training procedure, and preprocessing steps, are contributing to the enhancement of the model. Although, it is important to keep in mind that visual inspection alone might not provide a comprehensive evaluation of the model's performance. Quantitive metrics such as DSC, mIoU or MSE, could give a more precise assessment of models' performance.

Framework	Year	mIoU ↑	Se ↑
Chiu et al. [49]	2015	0.8765	0.8809
Roy et al. [50]	2017	0.8819	0.8954
Kepp et al. [51]	2019	0.8936	0.9089
Wei,Peng. [52]	2020	0.8976	0.9134
DeepRetina [36]	2020	0.9005	0.9229

Table 5.4: 9 layer segmentation SOTA Models

5.5 Overview of Segmentation Results

The DukePeople dataset seems to limit the performance of the models in this study due to inaccuracies in the annotated dataset. For instance, Figure 5.8 displays an annotated image of an individual impacted by AMD. This figure clearly displays degeneration of the eye's macula,

causing the neurological layers to take on a wave-like appearance. However, the ground truth annotated image (b), doesn't seem to follow that pattern.

Such inconsistencies can have a considerable impact on model performance, especially in tasks like image segmentation where accurate ground truth labels are key for successful training. It is highly possible that the performance of these models is being obstructed by the ground truth annotations in cases of images affected by AMD. This emphasises the critical need for careful and accurate data annotation, particularly in the context of medical image analysis where inaccuracies can have severe consequences.

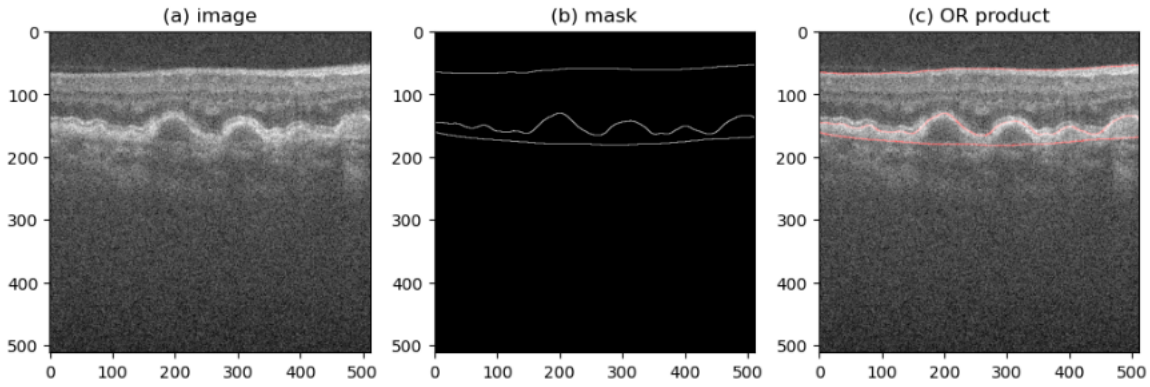

Figure 5.8: Incorrect annotations of DukePeople

5.6 Quantitative experiment on UKBioBank

This experiment evaluates the performance of the best-proposed model, TransUNet, on an image from UKBioBank. A three-layer segmentation will be generated to delineate the different neurological structures, followed by the image reconstruction part and finally, quantitive evaluation of the neural structure (See Figure 4.1).

Figure 5.9 demonstrates the effectiveness of the TransUNet model, trained in this study, when applied to an image from the UK BioBank. The reconstructed image, interestingly, reveals that despite the fact that the model was not trained specifically on images from the UK BioBank, which were probably taken by different imaging (OCT) devices, it still successfully delineates the structure.

After inputting this image into the walking pixel algorithm, it reveals that the total pixel count between the two layers (3 boundaries) stands at 33441, and 3156, respectively. Research

claims [46] that OCT images in UKbiobank were captured using a raster scan protocol measuring $6mm \cdot 6mm$. From this information, we can deduce that the thickness of the layers is $4592.42\mu m^2$ and $433.41\mu m^2$

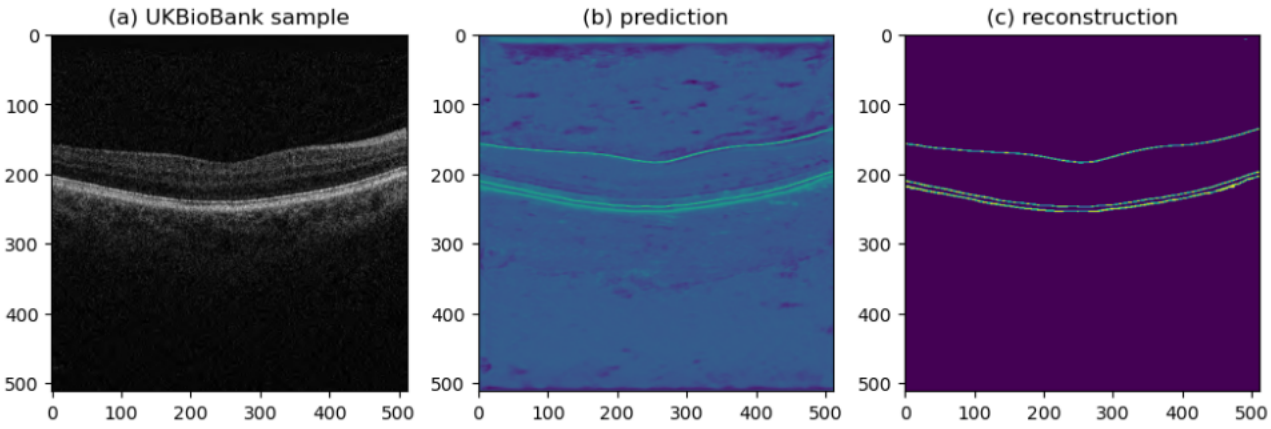

Figure 5.9: Sample of prediction using UKBioBank

Another example of segmentation of a sample taken from UK BioBank is depicted in Figure 5.10. Although the INL layer has been predicted successfully, considering the slight deformation, the area around the brunch's membrane wasn't clearly visible, thus the model struggled to predict its retinal layer. Therefore, the quantitative algorithm cannot be applied to such data. This is most likely due to the fact that the models were not fine tunned on UKBioBank, but only on DukePeople dataset.

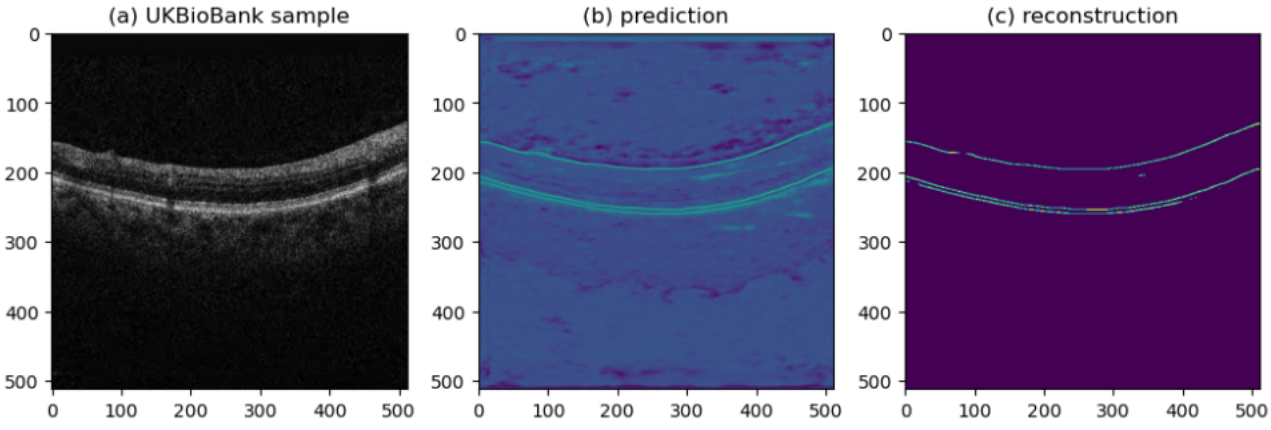

Figure 5.10: Sample of a partially predicted image from UKBioBank

5.7 Explainable AI

Explainable AI, also known as Interpretable AI, is a branch of research that seeks to provide clear, understandable explanations for the decisions behind neural networks. This section will try to bring closer the meaning and interpretability of the AI models used in this thesis for segmenting retinal images

While it is true that deep learning models, can be difficult and challenging to understand to the bone, one of the major criticisms is that they act as 'black boxes', which however is not entirely accurate. The field of image recognition, including image segmentation, is an example of a problem which could provide clear and understandable explanations behind its decisions.

Since the introduction of the AlexNet model and the ImageNet database in 2012, CNNs and image segmentation have become deeply interconnected. Even with the proposal of transformers, there is a variation of visual transformers which implements the CNN encoders for describing images in the latent space. Essentially, CNN with the use of their convolutional maps, also known as feature maps, are able to create patterns for delineating important aspects of the image.

Generally speaking, CNN consists of multiple convolutional layers with a depth *'c'*, pulling layers, and activation functions. Each convolutional layer represents some patterns that the neural network holds for that particular domain of the problem.

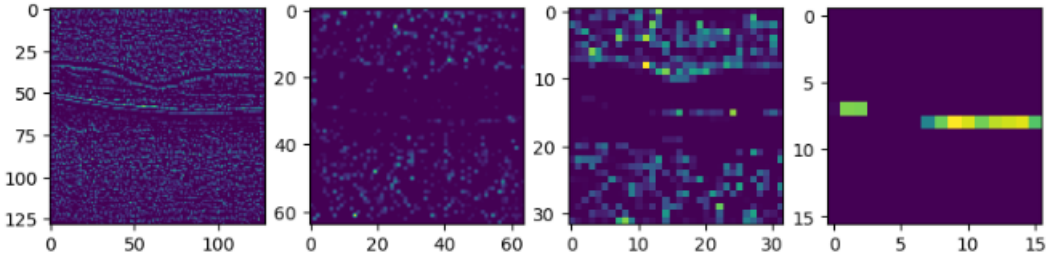

Figure 5.11: Resnet convolutions

Figure 5.11 depicts the depth of the convolutional layers, of the ResNetUNet trained in this study, and its depth over the first 4 layers. The depth refers to the number of filters used in each convolutional layer, each of which captures different features or patterns in the image.

This example is observing the filter of depth 26. In other words, it refers to the 26th

filter in the sequence of the convolutional layers. The first image in this figure represents the pattern created for the highest dimensional, and it could be interpreted as the input layer of the encoder. This layer is still somewhat self-explainable to us humans, as clear patterns of neurological boundaries are created. As we progress deeper in the network (left to right), these filters are becoming more complex and abstract, capturing higher-level features in the data.

When the displayed convolutions are applied to the image, the results would resemble what is depicted in Figure 5.12. This figure effectively highlights the areas of the image that the convolutional process is primarily focusing on. See Appendix C for more samples.

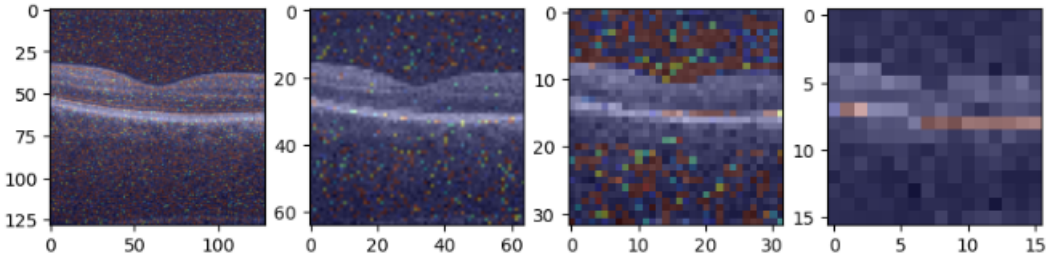

Figure 5.12: Resnet convolutions applied on an image

Ultimately, after generating low-level abstract features, also known as latent space information, those layers, including feature maps, are introduced into a visual transformer. This is a crucial step of the TransUNet network, as the visual transformer utilises this rich, abstract information, given by the ResNetUNet to make further processing and interpretation of the image. Similarly to the examples above, Figure 5.13 demonstrates the attention map of the visual transformer used in this work. In other words, these images demonstrate what the visual transformer 'sees' when it looks at the images of OCT.

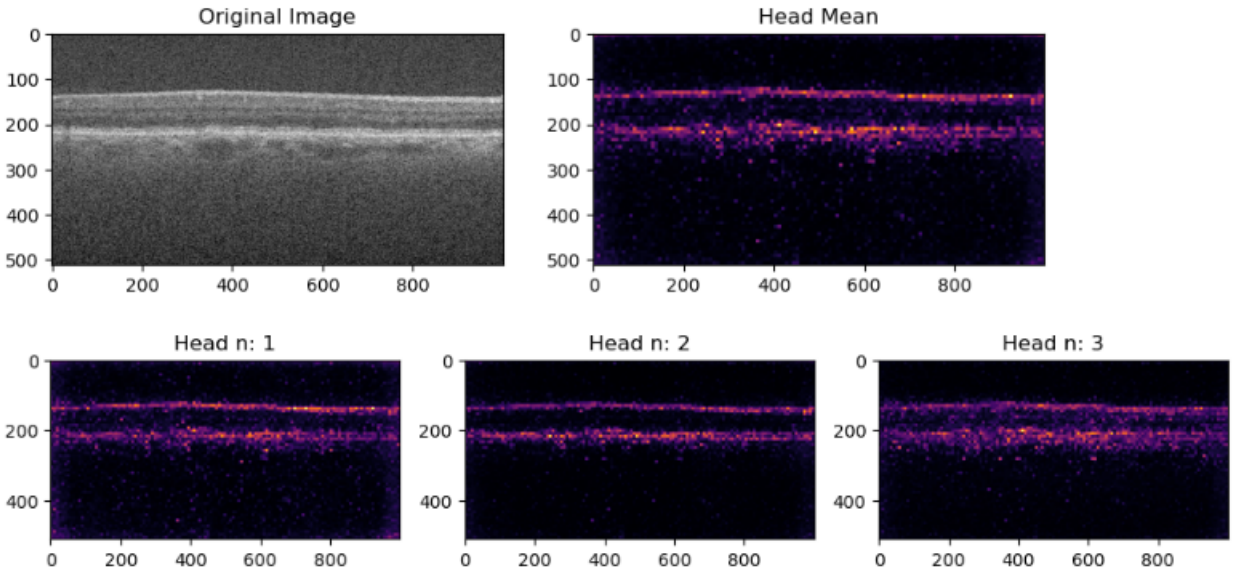

Figure 5.13: Visualising ViT head attention

To conclude, Figure 5.14 demonstrates the activation map of a convolutional layer situated in the decoder part of the TransUNet model. This image demonstrates where the network is focusing its attention.

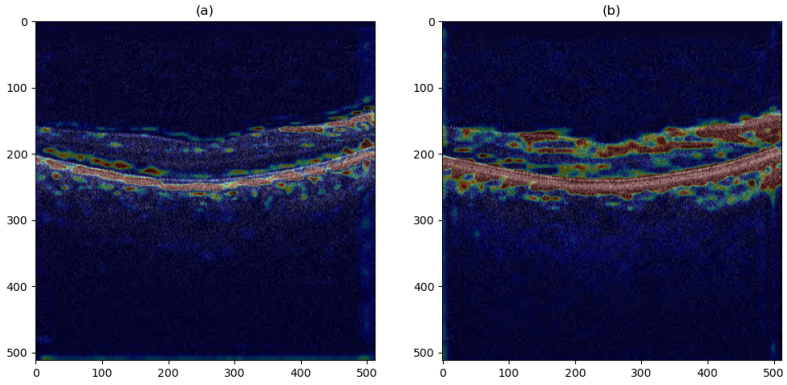

Figure 5.14: Class Attention Map of a convolutional layer in TransUNet network

Chapter 6

Conclusion and Future Directions

In summary, despite not achieving SOTA accuracy, this project has yielded successful results, despite the numerous challenges encountered during the research and development phase, which however required significant time and resource investments. Nonetheless, the encouraging outcomes derived from this project have demonstrated that the efforts and resources committed were well-spent and justified.

Over the course of this project, it was evident that the models do not derive significant benefits from the application of noise removal techniques. This may seem counterintuitive, as such tools are often used to improve the clarity and quality of the image, although these alterations, in most cases, are significant for human vision, they evidently do not impact convolution maps in the same way.

By implementing and evaluating both pre-trained and trained from scratch CNN architectures for the segmentation, it has been shown that convolutional neural networks remain the most used architecture in computer vision. This research has further underscored the exceptional capabilities of both CNN and Vision Transformers (ViT) in the field of computer vision and image segmentation. Moreover, the project has reaffirmed the status of UNet and its variants in medical image segmentation. Despite the emergence of newer technologies, such as transformers, UNet continues to hold its ground as a crucial building block for such tasks.

An interesting observation was made during the training process of a pre-trained backbone ResNet101 of the ResNetUNet model which reached its peak accuracy on the first epoch. This outcome proved the efficacy of pre-trained models, which not only reduce the time required

for training but also, conserve computational resources. These results essentially highlight the benefit of transfer learning, providing compelling evidence for its continued use in future projects.

The performance of the ResNetUNet model in contrast to the TransUNet model, which took 1 and 28 hours to compute, respectively, demonstrates that although the pre-defined weights are crucial for achieving early high accuracy, training from scratch could still yield superior results.

Class Activation Maps (CAM) were employed as discussed in Section 5.7 to address the decisions and reasoning behind models' decision-making process. Moreover, the visualisation of the multi-head attention mechanism was displayed too. By visualising which part of the image a network is giving attention to when making predictions, CAMs can help understand the reasoning behind a model's output. Moreover, such tools are particularly useful in models used for classification with healthcare applications where interpretability is crucial for defining new biomarkers.

This work has demonstrated an emerging need for an alternative, more robust and accurately annotated dataset, in order to achieve better performance and SOTA algorithms. Furthermore, better annotated data would increase the reliability of such models, allowing their applicability in medical diagnosis. Just as important, which was observed during the research, there is an urgent need for greater sharing of tools and resources within the research. All of the SOTA algorithms presented in the relevant work own an annotated dataset for training and testing purposes, which however is not shared within the research. Providing such a dataset would potentially yield newer SOTA algorithms and findings in this field.

Looking forward, an exciting opportunity lies in enriching the small-scale datasets, such as Duke People Chiu [49] which comprises around 100 9-layer segmented images. This could be achieved through the application of recent breakthroughs and SOTA algorithms, diffusion models, in generating synthetic data, thereby decreasing the need for large-scale human-annotated datasets.

Considering the lack of annotated data, exploring options beyond supervised learning could be beneficial. For instance, Singular spectrum decomposition, a time-series methodology, has the potential to identify patterns in pixel changes across the layers of neurological structures.

Although many difficulties and challenges were experienced during the course of this project, a wealth of knowledge has been obtained as well as broad and worthy experience in implementing SOTA architectures has been gained. Finally, the project achieved its ultimate objective by

publishing the project as an open-source project on a personal GitHub repository. Therefore, all of the aims and goals initially proposed were successfully accomplished, surpassing the initial expectations.

Chapter 7

Statement of Ethics

The following ethical considerations were carried out to ensure that the project was executed in a manner that was both ethical and lawful. In relation to the project, measures were taken regarding data management to guarantee the confidentiality of the data. Moreover, the project was assessed for compliance with a number of relevant regulations and statuses, including the Data Protection Act of 1990, the Computer Misuse Act (CMA) and the Code of Conduct.

7.1 Legal Considerations

7.1.1 Informed Consent

This project did not involve the active participation of human subjects or the acquisition of data. Thus the traditional need for informed consent and permission did not apply.Altough it wasn't necessary, legal considerations were still respected. For instance, in the case of the DukePeople dataset, the third-party dataset, the images were sourced from repositories where participants had given consent for their data to be used in research [44].

Additionally, UKBioBank is also third-party data which was managed by the university research. That is fully anonymised and was initially collected from participants around the UK who provided their full consent [53].

7.1.2 Data Confidentiality

Even though this research did not involve direct data collection from human subjects, data confidentiality was still respected. All the data collected from third-party sources were anonymised and de-identified. There was no additional information acquired from the datasets but the images used for training the models.

The UK BioBank data was managed by the university, and data security measures were put in place to prevent unauthorized access.

7.1.3 Intellectual Property

All software tools employed in this project are open-source, and any code that was used or modified falls under the MIT License. The license allows for commercial use, modification, distribution and private use [54].

7.1.4 Copyright

Given that this project and the final report are examined by Turnitin's verification software, any attempt by an individual to plagiarize this research and present it as their own would be detected by this system [55].

7.1.5 Computer Misuse Act

During the course of this project, an effort was made to adhere to legal regulations, ensuring that the functionality of any code was strictly confined to its intended purpose. Furthermore, there is no hidden functionality in the code of this project [56].

7.2 Ethical Considerations

7.2.1 Do Not Harm

This project serves as a proof of concept to demonstrate the potential of using a combination of visual transformers and convolutional neural network architecture for the segmentation of the neural structure in the human retina. However, it is worth mentioning that the accuracy

of the models achieved in this project does not reach a level that would be acceptable for clinical applications. These models were explicitly designed and created for research purposes. Therefore, no person should take the output of the models created as medical advice.

7.2.2 Data Protection Act

All the data used in this project was stored securely on the university hardware and all the computations were acquired on the university clusters. This measure was put in place to ensure compliance with the Data Protection Act, which mandates that data must be "handled in a way that ensures appropriate security, including protection against unlawful or unauthorised processing, access, loss, destruction or damage"

After the completion of the project, all the data acquired by the third-party sources will be deleted, in accordance with the Data Protection Act's "data must be kept for no longer than is necessary" [57].

7.2.3 Social Responsibility

The research undertaken in this project aims to contribute to the field of studying medical imaging, as well as to the field of neuroscience exploring neurodegenerative diseases. This project could be further used for early disease detection, which is crucial for effective treatment and early disease detection. The results conducted by this research display potential use in research for detecting necessary biomarkers for the early detection of neurodegenerative disease.

Appendix A

Data Augmentation

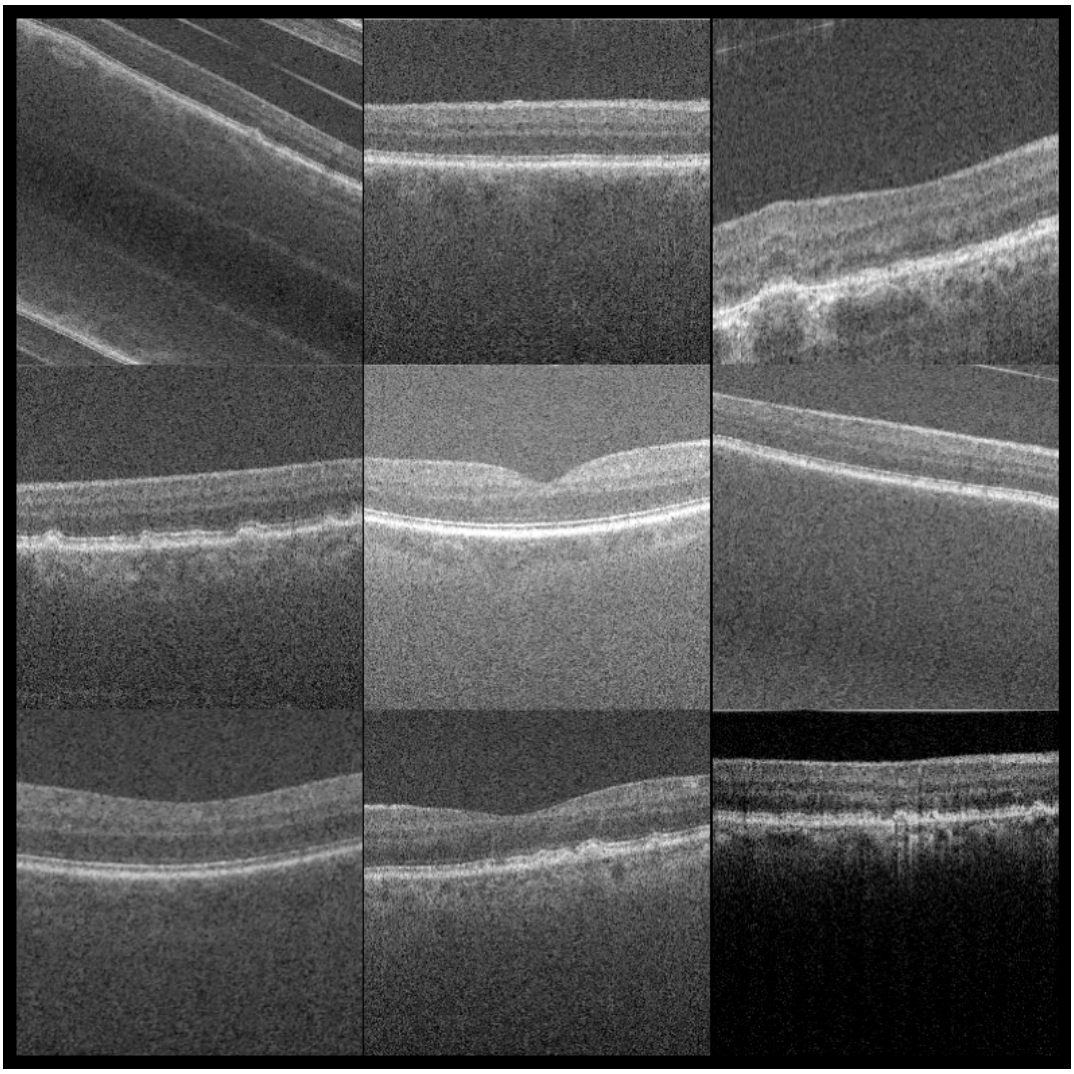

Figure A.1: Data augmentation of ground images of train set

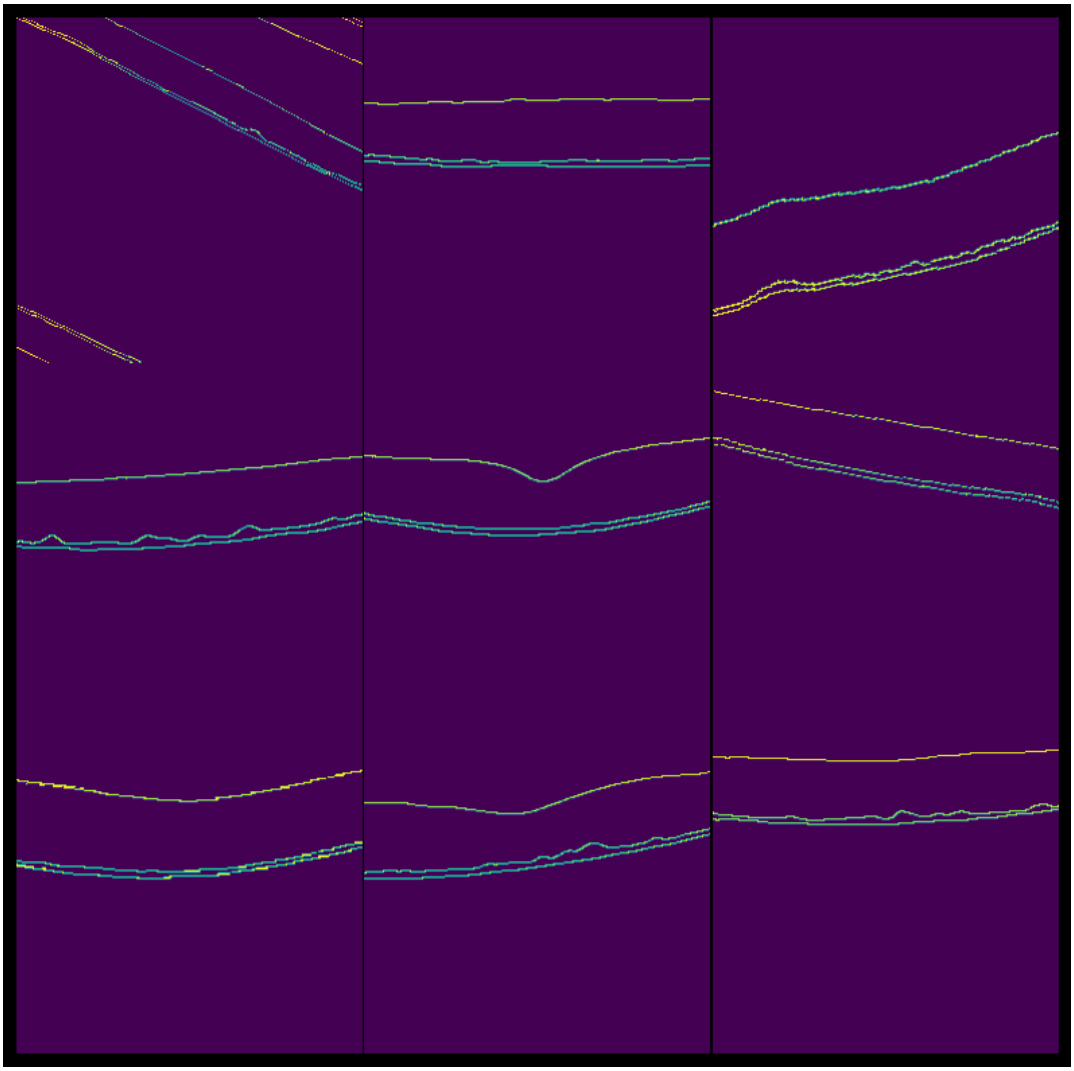

Figure A.2: Data augmentation of mask images of train set

Appendix B

Duke People Dataset

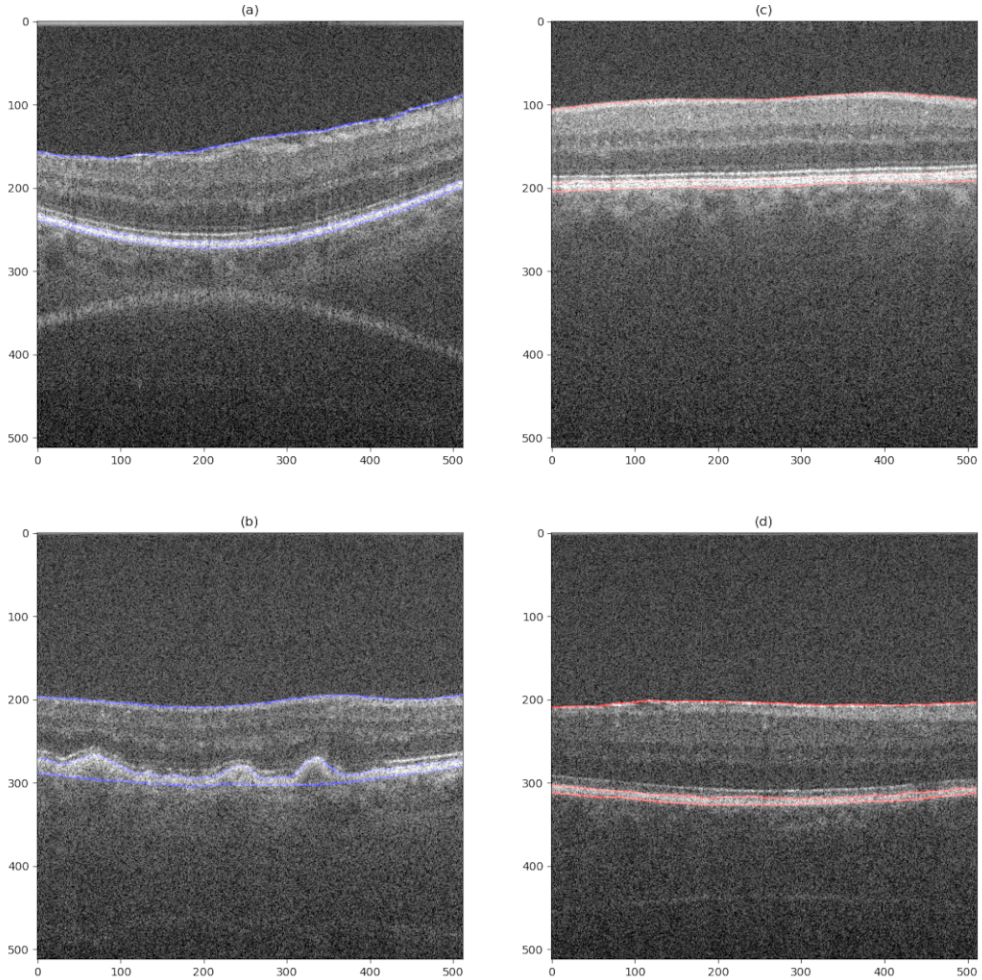

Figure B.1: AMD (a),(b); Controlled subjects (c),(d)

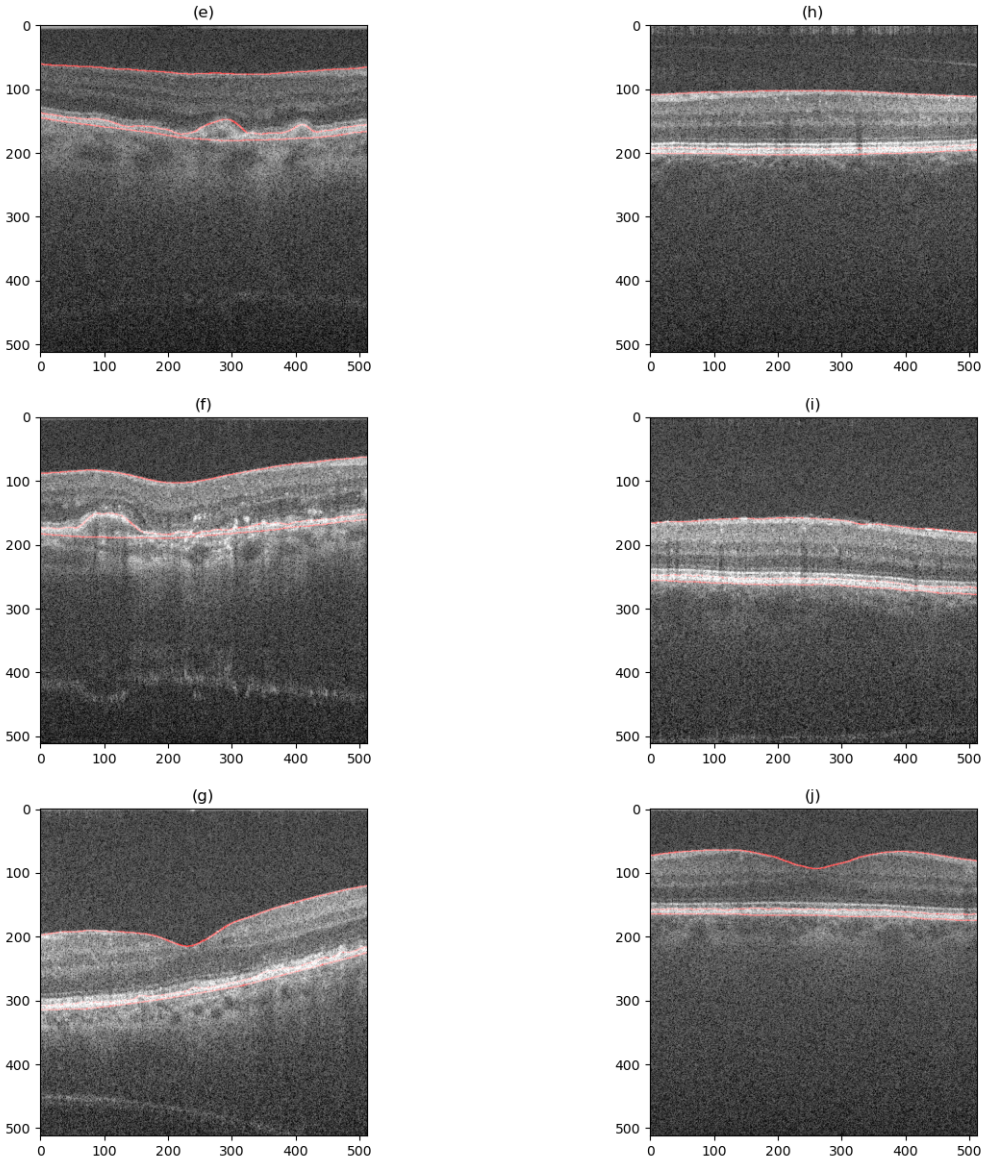

Figure B.2: AMD (a),(b); Controlled subjects (c),(d)

Appendix C

Explainable AI

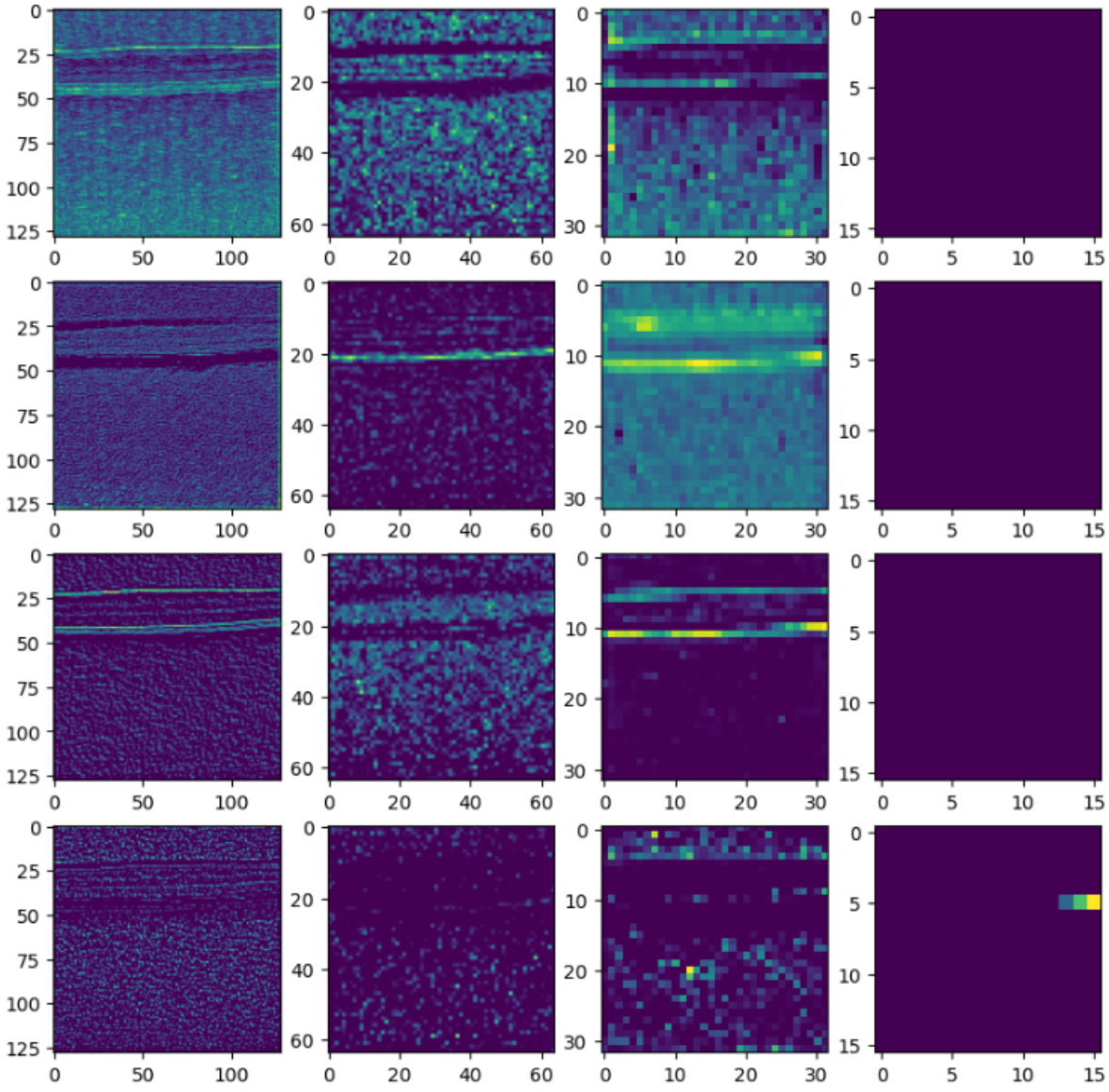

Figure C.1: 0th,4th,20th,26th dimension for layers 1,2,3 and 4 of the ResnetUNet base layer

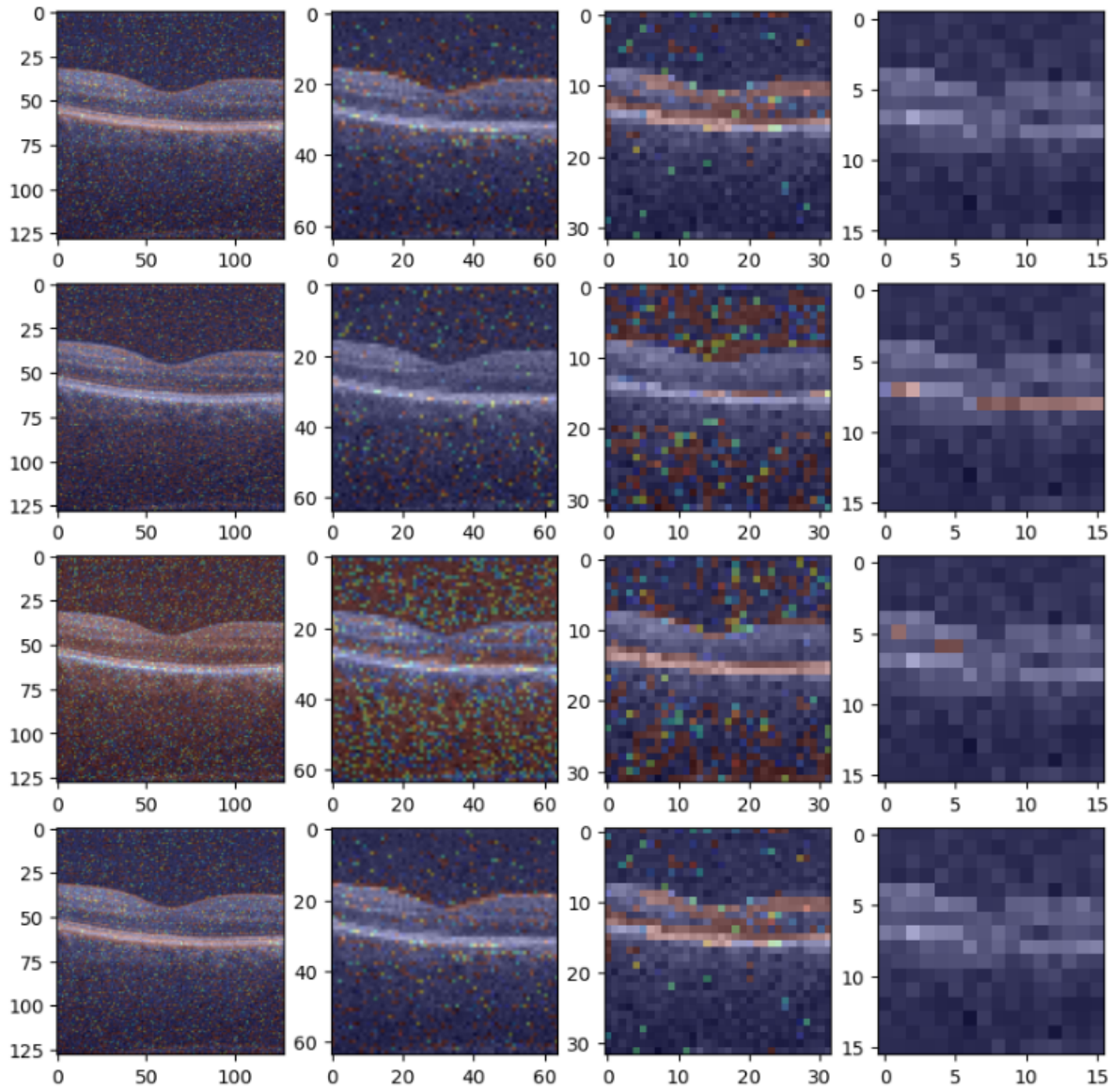

Figure C.2: Applied feature maps of ResNetUNet for layers 1,2,3,4 depth 2,4,26,28

Bibliography

[1] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, and M. P. Lungren, "Preparing medical imaging data for machine learning," *Radiology*, vol. 295, no. 1, pp. 4–15, 2020.

[2] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, 2020. [Online]. Available: https://www.mdpi.com/2078-2489/11/2/125

[3] A. Association, *2022 Alzheimer's Disease Facts and Figures*, 2023.

[4] S. Wang, H.-Y. Liu, Y.-C. Cheng, and C.-H. Su, "Exercise dosage in reducing the risk of dementia development: Mode, duration, and intensity—a narrative review," *International journal of environmental research and public health*, vol. 18, no. 24, p. 13331, 2021.

[5] G. D. Hildebrand and A. R. Fielder, "Anatomy and physiology of the retina," 2011.

[6] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Puliafito, and J. G. Fujimoto, "Optical coherence tomography," *Science*, vol. 254, no. 5035, pp. 1178–1181, 1991. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1957169

[7] P. A. Keane, S. Liakopoulos, R. V. Jivrajka, K. T. Chang, T. Alasil, A. C. Walsh, and S. R. Sadda, "Evaluation of Optical Coherence Tomography Retinal Thickness Parameters for Use in Clinical Trials for Neovascular Age-Related Macular Degeneration," *Investigative Ophthalmology Visual Science*, vol. 50, no. 7, pp. 3378–3385, 07 2009. [Online]. Available: https://doi.org/10.1167/iovs.08-2728

[8] J. C. Bavinger, G. E. Dunbar, M. S. Stem, T. S. Blachley, L. Kwark, S. Farsiu, G. R. Jackson, and T. W. Gardner, "The Effects of Diabetic Retinopathy and Pan-Retinal

Photocoagulation on Photoreceptor Cell Function as Assessed by Dark Adaptometry," *Investigative Ophthalmology Visual Science*, vol. 57, no. 1, pp. 208–217, 01 2016. [Online]. Available: https://doi.org/10.1167/iovs.15-17281

[9] C. A. Puliafito, M. R. Hee, C. P. Lin, E. Reichel, J. S. Schuman, J. S. Duker, J. A. Izatt, E. A. Swanson, and J. G. Fujimoto, "Imaging of macular diseases with optical coherence tomography," *Ophthalmology*, vol. 102, no. 2, pp. 217–229, 1995. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0161642095310329

[10] B. Knoll, J. Simonett, N. J. Volpe, S. Farsiu, M. Ward, A. Rademaker, S. Weintraub, and A. A. Fawzi, "Retinal nerve fiber layer thickness in amnestic mild cognitive impairment: Case-control study and meta-analysis," *Alzheimer's Dementia: Diagnosis, Assessment Disease Monitoring*, vol. 4, pp. 85–93, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352872916300409

[11] R. S. Maldonado, P. Mettu, M. El-Dairi, and M. T. Bhatti, "The application of optical coherence tomography in neurologic diseases," *Neurology: Clinical Practice*, vol. 5, no. 5, pp. 460–469, 2015.

[12] G. Song, E. T. Jelly, K. K. Chu, W. Y. Kendall, and A. Wax, "A review of low-cost and portable optical coherence tomography," *Progress in Biomedical Engineering*, vol. 3, no. 3, p. 032002, may 2021. [Online]. Available: https://dx.doi.org/10.1088/2516-1091/abfeb7

[13] D. Varghese, R. Bauer, D. Baxter-Beard, S. Muggleton, and A. Tamaddoni-Nezhad, "Human-like rule learning from images using one-shot hypothesis derivation," in *Inductive Logic Programming: 30th International Conference, ILP 2021, Virtual Event, October 25–27, 2021, Proceedings*. Springer, 2022, pp. 234–250.

[14] S. J. Chiu, X. T. Li, P. Nicholas, C. A. Toth, J. A. Izatt, and S. Farsiu, "Automatic segmentation of seven retinal layers in sdoct images congruent with expert manual segmentation," *Optics express*, vol. 18, no. 18, pp. 19 413–19 428, 2010.

[15] Information and P. C. of Ontario, "De-identification guidelines for structured data," no. 1, pp. 4–15, 2016.

[16] M. M. C. RSNA. (2019) Ctpthe rsna clinical trial processor. [Online]. Available: https://mircwiki.rsna.org/index.php?title=MIRC_CTP

[17] K. El Emam and F. K. Dankar, "Protecting privacy using k-anonymity," *Journal of the American Medical Informatics Association*, vol. 15, no. 5, pp. 627–637, 2008.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[19] S. Chawla, P. Nakov, A. Ali, W. Hall, I. Khalil, X. Ma, H. Taha Sencar, I. Weber, M. Wooldridge, and T. Yu, "Ten years after imagenet: A 360° perspective on ai," *arXiv e-prints*, pp. arXiv–2210, 2022.

[20] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into imaging*, vol. 9, pp. 611–629, 2018.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

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

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

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[25] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[26] M. S. Anat London, Inbal Benharm, "The retina as a window to the brain—from eye research to cns disorders," *Nature Reviews Neurology*, Jan. 2013.

[27] L. Fang, D. Cunefare, C. Wang, R. H. Guymer, S. Li, and S. Farsiu, "Automatic segmentation of nine retinal layer boundaries in oct images of non-exudative amd patients using deep learning and graph search." *Biomedical optics express*, vol. 8 5, pp. 2732–2744, 2017.

[28] D. Koozekanani, K. Boyer, and C. Roberts, "Retinal thickness measurements from optical coherence tomography using a markov boundary model," *IEEE transactions on medical imaging*, vol. 20, no. 9, pp. 900–916, 2001.

[29] J. Oliveira, S. Pereira, L. Gonçalves, M. Ferreira, and C. A. Silva, "Multi-surface segmentation of oct images with amd using sparse high order potentials," *Biomedical optics express*, vol. 8, no. 1, pp. 281–297, 2017.

[30] X. Chen, M. Niemeijer, L. Zhang, K. Lee, M. D. Abràmoff, and M. Sonka, "Three-dimensional segmentation of fluid-associated abnormalities in retinal oct: probability constrained graph-search-graph-cut," *IEEE transactions on medical imaging*, vol. 31, no. 8, pp. 1521–1531, 2012.

[31] P. Gholami, P. Roy, M. K. Parthasarathy, A. Ommani, J. Zelek, and V. Lakshminarayanan, "Intra-retinal segmentation of optical coherence tomography images using active contours with a dynamic programming initializatiopekala2019103445n and an adaptive weighting strategy," in *Optical Coherence Tomography and Coherence Domain Optical Methods in Biomedicine XXII*, vol. 10483. SPIE, 2018, pp. 61–66.

[32] M. Pekala, N. Joshi, T. A. Liu, N. Bressler, D. C. DeBuc, and P. Burlina, "Deep learning based retinal oct segmentation," *Computers in Biology and Medicine*, vol. 114, p. 103445, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482519303221

[33] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 11–19.

[34] P. Diederik and B. Jimmy, "Kingma, adam: A method for stochastic optimization," 2015.

[35] J. Tian, B. Varga, E. Tatrai, P. Fanni, G. M. Somfai, W. E. Smiddy, and D. C. Debuc, "Performance evaluation of automated segmentation software on optical coherence tomography volume data," *Journal of Biophotonics*, vol. 9, no. 5, pp. 478–489, 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/jbio.201500239

[36] Q. Li, S. Li, Z. He, H. Guan, R. Chen, Y. Xu, T. Wang, S. Qi, J. Mei, and W. Wang, "Deepretina: layer segmentation of retina in oct images using deep learning," *Translational Vision Science & Technology*, vol. 9, no. 2, pp. 61–61, 2020.

[37] J. Kugelman, D. Alonso-Caneiro, S. A. Read, J. Hamwood, S. J. Vincent, F. K. Chen, and M. J. Collins, "Automatic choroidal segmentation in oct images using supervised deep learning methods," *Scientific reports*, vol. 9, no. 1, pp. 1–13, 2019.

[38] J. Hamwood, D. Alonso-Caneiro, S. A. Read, S. J. Vincent, and M. J. Collins, "Effect of patch size and network architecture on a convolutional neural network approach for automatic segmentation of oct retinal layers," *Biomedical optics express*, vol. 9, no. 7, pp. 3049–3066, 2018.

[39] J. A. Sousa, A. Paiva, A. Silva, J. D. Almeida, G. Braz Junior, J. O. Diniz, W. K. Figueredo, and M. Gattass, "Automatic segmentation of retinal layers in oct images with intermediate age-related macular degeneration using u-net and dexined," *Plos one*, vol. 16, no. 5, p. e0251591, 2021.

[40] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, 1998, pp. 839–846.

[41] M. Mujat, R. C. Chan, B. Cense, B. H. Park, C. Joo, T. Akkin, T. C. Chen, and J. F. De Boer, "Retinal nerve fiber layer thickness map determined from optical coherence tomography images," *Optics Express*, vol. 13, no. 23, pp. 9480–9491, 2005.

[42] D. C. Fernández, H. M. Salinas, and C. A. Puliafito, "Automated detection of retinal layer structures on optical coherence tomography images," *Optics express*, vol. 13, no. 25, pp. 10 200–10 216, 2005.

[43] E. Stevens, L. Antiga, and T. Viehmann, *Deep learning with PyTorch.* Manning Publications, 2020.

[44] S. Farsiu, S. J. Chiu, R. V. O'Connell, F. A. Folgar, E. Yuan, J. A. Izatt, C. A. Toth, A.-R. E. D. S. . A. S. D. O. C. T. S. Group *et al.*, "Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography," *Ophthalmology*, vol. 121, no. 1, pp. 162–172, 2014.

[45] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray *et al.*, "Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS medicine*, vol. 12, no. 3, p. e1001779, 2015.

[46] P. A. Keane, C. M. Grossi, P. J. Foster, Q. Yang, C. A. Reisman, K. Chan, T. Peto, D. Thomas, P. J. Patel, and U. B. E. V. Consortium, "Optical coherence tomography in the uk biobank study–rapid automated analysis of retinal thickness for large population-based studies," *PLoS One*, vol. 11, no. 10, p. e0164095, 2016.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[48] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial intelligence and machine learning for multi-domain operations applications*, vol. 11006. SPIE, 2019, pp. 369–386.

[49] S. J. Chiu, M. J. Allingham, P. S. Mettu, S. W. Cousins, J. A. Izatt, and S. Farsiu, "Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema," *Biomedical optics express*, vol. 6, no. 4, pp. 1172–1194, 2015.

[50] A. G. Roy, S. Conjeti, S. P. K. Karri, D. Sheet, A. Katouzian, C. Wachinger, and N. Navab, "Relaynet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," *Biomedical optics express*, vol. 8, no. 8, pp. 3627–3642, 2017.

[51] T. Kepp, J. Ehrhardt, M. P. Heinrich, G. Hüttmann, and H. Handels, "Topology-preserving shape-based regression of retinal layers in oct image data using convolutional neural networks," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 1437–1440.

[52] H. Wei and P. Peng, "The segmentation of retinal layer and fluid in sd-oct images using mutex dice loss based fully convolutional networks," *Ieee Access*, vol. 8, pp. 60 929–60 939, 2020.

[53] "UK BioBank," https://www.ukbiobank.ac.uk/, 2008, [Online; accessed 16-May-2023].

[54] P. of the United Kingdom, ""the mit license,"," "https://opensource.org/licenses/MIT, 2023, [Online; accessed 16-May-2023].

[55] "TurnitIn," https://www.turnitin.com/regions/uk., 2023, [Online; accessed 16-May-2023].

[56] P. of the United Kingdom, "Computer misuse act 1990," https://www.legislation.gov.uk/ukpga/1990/18/contents, 1990, [Online; accessed 16-May-2023].

[57] ——, "Data Protection Act 2018," ttps://www.legislation.gov.uk/ukpga/2018/12/contents, 2018, [Online; accessed 16-May-2023].